UK Biobank

Biomarker assay quality procedures: approaches used to minimise systematic and random errors (and the wider epidemiological implications)

Version 1.2

http://www.ukbiobank.ac.uk/

02 April 2019

Contents

1.0 Overview 3
2.0 Sample collection and processing 3
2.1 Assays performed immediately 4
3.0 Biochemistry assay selection 4
4.0 Sample retrieval for the biochemistry assays 4
5.0 Quality performance checks 5
5.1 Quality system and scope of laboratory accreditation 5
5.2 Internal and external quality control procedures 5
6.0 Identification of issues during internal quality control checks 5
6.1 Invalid results 6
6.2 Dilution issue 6
6.3 Identification of laboratory drift of assay results 7
7.0 Proposed solutions and epidemiological considerations 7
7.1 Dataset provided in showcase 7
7.2 Sensitivity analyses 8
7.3 Use of repeated measures 8
8.0 Future considerations 8

1.0 Overview

In order to enhance further the value of the UK Biobank resource to researchers, UK Biobank has embarked on a project to measure a wide range of biochemical markers in biological samples collected at baseline (2006-2010) in all participants with an available sample. This comprised about 480,000 participant samples from the recruitment visit and about 18,000 samples collected at a repeat assessment (2012-2013).

This document describes the approach that UK Biobank has taken to minimise and mitigate the effects of error (both systematic bias and random error) that such data are liable to include, in order to provide high-quality biomarker data. This includes our approach to sample collection and processing, sample retrieval, and monitoring of assay data in order to minimise drift, bias and measurement error. We also describe any issues that have been identified and our proposed solutions.

2.0 Sample collection and processing

During recruitment of 500,000 participants into UK Biobank, a series of biological samples were collected comprising blood (about 45 ml), urine (about 9 ml) and, for the last 100,000 participants, saliva. The samples were collected in different collection vessels so that a variety of preservatives, anti-coagulants and clot accelerators could be used to allow the widest possible range of assays that could plausibly be envisaged for the future (Table 1).

The collection vessels (vacutainers and collection pots) were then processed on a variety of automation systems to create, for some sample types, multiple aliquots for long-term storage. Half of sample aliquots are stored in a fully automated -80°C working archive (1) and half in a manual, nitrogen-vapour back-up archive located at separate sites to protect them from degradation caused by freeze-thawing, or loss due to breakdown of a single archive site (Table 1).

Sample collection tube	Fractions	Number of aliquots	
		-80°C	Liquid nitrogen
EDTA x2	Plasma	6	2
	Buffy coat	1	1
	Red cells	-	2
Lithium heparin (PST)	Plasma	3	1
Silica clot accelerator (SST)	Serum	3	1
Acid citrate dextrose	DMSO blood	-	2
EDTA	Haematology	-	-
Urine	Urine	4	2
Tempus tube (RNA)	Whole blood	-	6
Saliva	Mixed saliva	-	2
Total		**17**	**19**

Table 1: Sample collection and maximum number of aliquots created for each sample type

Abbreviations: EDTA: Ethylenediaminetetraacetic acid; PST, plasma separator tube; SST, serum separator tube.

The sample handling procedures (2) were the result of extensive consultation and piloting to try and ensure that the proposed procedures were fit for purpose and feasible at scale (3). For example, the pilot studies showed that a very wide range of assays could be performed in whole blood and urine samples maintained at 4°C for up to 36 hours prior to processing and

storage (2, 4). As such, the samples were minimally processed at the assessment centres, with most of the processing conducted at the central laboratory using more efficient and reliable automated systems. The only processing that was done immediately at the assessment centre involved inverting the plasma and serum tubes to mix the preservative/anti-coagulant with the blood and then allowing the serum tube to clot at room temp for 30 minutes. Both tubes contained a gel plug that formed a barrier to cellular material while allowing the plasma/serum to pass through during centrifugation (at 4°C), thereby producing sample separation. All tubes were refrigerated (with the exception of the acid citrate dextrose tube which was held at room temperature) until the end of the day when they were packed (with temperature logging devices) and transported to UK Biobank's central processing and archiving facility in Stockport.

At the central laboratory, all samples were predominantly processed using custom-designed industrial-scale automation systems to generate about 25,000 sample aliquots per day (resulting in 15 million 1.4 ml aliquots for the full cohort). A small proportion of samples were manually aliquoted. The extensive use of automation ensured that all of the samples were processed quickly, with an average time of 24 ± 2.5 hours between venepuncture and sample storage. This was achieved by ensuring that the samples were processed at the central facility in the same chronological order in which they were collected. The use of automation also allowed for a carefully controlled data trail linking each aliquot correctly to the participant from whom they derived (2).

2.1 Assays performed immediately

Standard haematological tests were performed on fresh whole blood within 24 hours of blood collection for all of the participants, and are available for researchers to access through the Data Showcase (https:/biobank.ndph.ox.ac.uk/showcase/label.cgi?id=100081).

3.0 Biochemistry assay selection

Overall, 36 biochemistry markers were selected for assay in all 500,000 participants, full details of which can be found here: https://www.ukbiobank.ac.uk/uk-biobank-biomarker-panel/. The biomarkers were selected for analysis because they represented established risk factors for disease, were established diagnostic measures, or characterised phenotypes not otherwise well assessed and were feasible to measure at scale. Overall, 29 biomarkers were measured in serum (SST tube), 4 in urine (plain tube), and 1 in red blood cells (EDTA tube; Table 1). The project was co-ordinated by the UK Biobank Enhancements Working Group, with input from external experts as required.

4.0 Sample retrieval for the biochemistry assays

Only those tubes that were required for the biochemistry assays were extracted from the freezer and thawed, with the rest of the aliquots on each plate returned still frozen to the working archive (in order to avoid unnecessary freeze-thaw events).

In order to avoid biases in epidemiological comparisons, we developed algorithms designed to select aliquots that avoided clustering of samples by geography, collection dates or time of day. Simulations of the performance of this picking strategy demonstrated its effectiveness and efficiency for the different sample types and ensured that participant samples were analysed in an effectively random manner during the project (please see Appendix 3 of the serum biochemistry companion document for further details: http://ace.ctsu.ox.ac.uk/showcase/refer.cgi?id=1227). This enabled us to better assess the variation caused by lab/assay drift (as the mean biomarker concentration across batches and analysers should effectively be the same).

5.0 Quality performance checks

We employed a series of robust and detailed quality procedures designed to minimise drift, bias and measurement uncertainty, details of which are provided below.

5.1 Quality system and scope of laboratory accreditation

Multiple immunoassay and clinical chemistry analysers were used to measure the biochemistry markers, details of which are provided in our online companion documents for serum (https://biobank.ndph.ox.ac.uk/showcase/docs/serum_biochemistry.pdf), urine (https://biobank.ndph.ox.ac.uk/showcase/docs/urine_assay.pdf) and red blood cells (https://biobank.ndph.ox.ac.uk/showcase/refer.cgi?id=2405), respectively. These analysers were designed according to the internationally recognised standard for testing and calibration laboratories (ISO 17025:2005). During the project, the UK Biobank laboratory was successfully externally audited against the ISO 17025:2005 standard on 27th Dec 2015 for the urine and HbA1c assays, and on 17th Oct 2016 for the serum assays.

5.2 Internal and external quality control procedures

Each assay was registered with an external quality assurance (EQA) scheme and assay performance was externally verified via the results returned from participation in these schemes. We also followed a rigorous internal quality control (QC) protocol to assess precision (using different concentrations of QC samples over multiple batches and analysers) and accuracy and bias (using EQA or other commercially validated material). We verified that the assays were linear over the observed reportable range (using commercial linearity standards and low concentration samples) and that there were no carryover effects (using low and high concentration samples analysed consecutively in a standardised sequence). We also assessed potential assay interferences on each sample that could cause falsely high or low results. Please refer to our online companion document on QC metrics for further details (https://biobank.ndph.ox.ac.uk/showcase/refer.cgi?id=1227).

6.0 Identification of issues during internal quality control checks

All of the participant and QC data generated were reviewed both during the project (to identify and address any issues in real time) and to allow retrospective adjustments to be made, where required.

6.1 Invalid results

We have excluded results from the Data Showcase where no data or an error was returned from the analyser, the values were outside the reportable range of the assay at the time of measurement or there was an aliquot problem. Overall, this affected 9% of assay results. In particular, oestradiol and rheumatoid factor had a high proportion of values below the lower reportable range (80% and 91%, respectively), which is to be expected given the age range of the UK Biobank population. Hence, researchers may wish to consider these values as 'naturally low' rather than 'missing' in order to maximise the scientific utility of these data. Further details of the reportable range for all assays are provided in Appendix 1 of the companion document.

6.2 Dilution issue

Shortly after the assay period started, the laboratory team discovered during routine quality control checks that, during the initial sample processing that occurred at the time of sample collection (for both the baseline and repeat assessment), some participant serum samples were inadvertently diluted during creation of the aliquots from the serum vacutainer. The dilution is believed to come from mixing of participant sample with system fluid (water) due to seals that failed to hold a system vacuum in the automated liquid handling systems.

At the end of the assay period, statistical models were fitted to estimate the magnitude of the unplanned dilutions at the time of aliquoting. A model incorporating shifts in dilutions over time of aliquoting provided an indication of the extent to which each sample was affected by dilution but did not provide an adequate means of correcting all results. Therefore, assay results have been corrected for the effects of differences in dilution associated with aliquot number only and some results for badly affected assays have been excluded on the basis of the estimated reduction in concentration of the sample. See the technical appendix for further details of the dilution issue and the statistical methods used to mitigate this issue.

Overall, the vast majority (92%) of assay results are not materially affected by dilution, with an estimated concentration of up to 1% lower (or higher) than that in an unaffected sample; 8.3% are diluted up to 10%; very few sample are diluted more than this. The dilution is systematic in that the magnitude of the dilution increases with increasing aliquot number (i.e. aliquot 1s are less affected than aliquot 2s, etc.). As soon as the issue was discovered, the lab prioritised use of aliquot 1 in order to reduce the impact of the dilution on the assay results. Hence, 98.5% of assay results from aliquot 1 (which accounts for 90% of participant serum samples) have an estimated concentration of up to 1% lower (or higher), and 1.5% have an estimated concentration up to 3% lower (or higher). Aliquot 3 is the most affected by dilution with all samples being diluted to some extent, although this only accounts for a small proportion (2%) of participant samples (Table 2). All results from aliquot 4 were excluded as there were not enough results for accurate assessment of the dilution problem, which appeared to be the most severe.

Assays with a naturally wide biological range are typically far less importantly affected from an epidemiological perspective by a given dilution, as small dilution errors are immaterial compared with their biological variation across the population. Conversely, assays with a

narrow biological range (e.g., calcium, total protein, phosphate and albumin) are more strongly affected.

As only one aliquot was dispensed for each of the EDTA sample tubes for red blood cells, the HbA1c assay results do not have a corresponding aliquot number associated with them.

		Estimated percentage reduction in sample concentration							
Aliquot Number	Number of samples	≤-2%	>-2% to ≤-1%	>-1% to <-1%	≥1% to <3%	≥3% to <5%	≥5% to <10%	≥10%[2]	Mean
Manual	9086	-	-	100%	-	-	-	-	-
1	418170	0.00%	0.21%	98.54%	1.25%	-	-	-	0.1%
2	29050	-	0.01%	20.55%	77.88%	1.56%	-	-	1.4%
3	10230	0.01%	-	-	0.02%	1.94%	97.97%	0.07%	6.4%
Total	466536	0.00%	0.19%	91.55%	5.97%	0.14%	2.15%	0.00%	0.3%

Table 2: Distribution of estimated percentage reductions in sample concentration by aliquot number (among baseline samples where the same aliquot was used for all assays) [1]

[1] The serum assays lipoprotein A, Gamma glutamyltransferase, C-reactive protein, rheumatoid factor and vitamin D were excluded from the analyses used to estimate the dilution factor, as was HbA1c (performed in red blood cells). Also excluded were participant samples that were processed into different aliquots for these assays (see technical appendix for further details).

[2] Results for serum assays with estimated percentage reduction in sample concentration ≥10% are set to missing in the Data Showcase data.

6.3 Identification of laboratory drift of assay results

Owing to the random selection of the participant samples and the sheer volume of results, it was possible to perform statistical analysis of the participant data to ensure that the day-to-day variation was within acceptable limits (as the overall mean of the biomarker levels should not vary across batches over time owing to the random plating of samples). These analyses showed that, with the power of large numbers, most assays had some detectable evidence of assay drift or jumps in values over time. As a result, we have corrected the majority of assay results for date-of-assay. See the technical appendix for further details of the methods used to correct for assay drift.

7.0 Proposed solutions and epidemiological considerations

7.1 Dataset provided in showcase

We have attempted to correct for unexpected dilution and laboratory drift, where appropriate, and the corrected assay results are available through the main Data Showcase. We have also included the estimated sample dilution factor and flags indicating the type of correction applied and reasons for exclusions. Please see the technical appendix for details of what data-fields are provided.

Please note: These corrections are a first-pass approach. As some assays are affected more than others, researchers should be cognisant of the size of the correction applied to their biomarkers of interest. For those researchers specifically interested in investigating the dilution issue more thoroughly, an extended dataset is available that includes all the original values

and lab parameters needed to develop other corrections, should they wish to do so. Further details of the data-fields available in the extended dataset are listed in the technical appendix. This extended dataset provides the opportunity for researchers to scrutinise the applicability of the correction across the whole of the assay range and potentially improve the correction and/or provide additional guidance on which samples to include for specific analyses.

7.2 Sensitivity analyses

91.6% of the assays for this project were performed either in aliquot 1 (because the laboratory team preferentially selected the least affected aliquot once the dilution issue was identified) or in manually aliquoted samples. Researchers could consider performing sensitivity analyses (such as comparing the results of analyses before and after restricting the dataset to assay results derived from aliquot 1) or stratifying the analyses by aliquot number to assess the potential impact of dilution (and its correction) on the interpretation of epidemiological results.

7.3 Use of repeated measures

UK Biobank was designed to collect repeat measures in random subsets of the cohort every few years in order to correct for sources of variation in the relative associations of baseline exposure measures with health outcomes (i.e. regression dilution). Because of fluctuations in the baseline measures, such associations often underestimate the strength of the real association between the "usual" (i.e., long-term average) risk factor value and disease risk (5). With respect to biomarkers, random variation in baseline values is caused by imprecise measurement (including laboratory/assay variation) and/or true short-term biological variability.

In UK Biobank, approximately 18,000 samples were collected at a repeat assessment approximately 5 years later (2012-2013) and were included in the biomarker project, principally to take into account regression dilution, which can be done using standard statistical approaches (5-7). However, because these samples were also processed using automated systems and hence were subjected to the same dilution effects as the baseline samples, they cannot be used to fully adjust for this specific source of error.

8.0 Future considerations

Because the extent of dilution differs across aliquots, UK Biobank will carefully consider the balance of aliquot numbers (i.e. either aliquot 1, 2, 3 or 4) to perform assays for future biomarkers. This will mitigate any systematic error associated with increasing amounts of dilution across aliquot number, although there will remain some variation in dilution across samples within a given aliquot number and potential bias in the correlations of the results with those in the main biochemistry panel. For those biomarker assays that are unlikely to be unduly affected by a small amount of error caused by dilution (e.g., semi-quantitative assays, assays with a naturally wide biological range or assays that require dilution prior to measurement), we will consider using higher-order aliquots (i.e. aliquot 2, 3 or 4). Conversely, for those assays known to be highly affected by dilution (i.e. with a low biological range), it would be more appropriate to use aliquot 1.

References

1. Owen JM, Woods P. Designing and implementing a large-scale automated -80 degrees C archive. *Int J Epidemiol* 2008; 37(Suppl 1):i56-61.

2. Elliott P, Peakman TC. The UK Biobank sample handling and storage protocol for the collection, processing and archiving of human blood and urine. *Int J Epidemiol* 2008; 37(2): 234-44.

3. Downey P, Peakman TC. Design and implementation of a high-throughput biological sample processing facility using modern manufacturing principles. *Int J Epidemiol* 2008; 37(Suppl 1):i46-50.

4. Jackson C, Best N, Elliott P. UK Biobank Pilot Study: stability of haematological and clinical chemistry analytes. *Int J Epidemiol* 2008; 37(Suppl 1):i16-22.

5. MacMahon S, Peto R, Cutler J, Collins R, Sorlie P, Neaton J, et al. Blood pressure, stroke, and coronary heart disease. Part 1: Prolonged differences in blood pressure: prospective observational studies corrected for the regression dilution bias. *Lancet* 1990; 335(8692):765-74.

6. Clarke R, Shipley M, Lewington S, Youngman L, Collins R, Marmot M, et al. Underestimation of risk associations due to regression dilution in long-term follow-up of prospective studies. *Am J Epidemiol* 1999; 150(4):341-53.

7. Lewington S, Thomsen T, Davidsen M, Sherliker P, Clarke R. Regression dilution bias in blood total and high-density lipoprotein cholesterol and blood pressure in the Glostrup and Framingham prospective studies. *J Cardiovasc Risk* 2003; 10(2):143-8.

Appendix: Biomarker Assays Quality Procedures: statistical investigation

Matthew Arnold[1] and Sarah Parish[2,1]

1 Clinical Trial Service Unit & Epidemiological Studies Unit, Nuffield Department of Population Health, University of Oxford

2 MRC Population Health Research Unit, Nuffield Department of Population Health, University of Oxford

1.0 Background

As described in the main Biomarker Assay Quality Procedures document, sample aliquots suffered from some unplanned dilutions, and in addition assay results suffered from some anticipated assay drift. The overarching aim of this investigation is to identify adequately robust data that can be released to Data Showcase. Investigation and consideration of anomalies is partial but further investigation will be most efficaciously addressed by also making the whole data available (as an extended dataset available under the User returned dataset section) to the UK Biobank researchers. This provides the opportunity for researchers to either use the Data Showcase values accepting that they may have (as yet uncovered) limitations, or for those with detailed knowledge of particular assays, access to corroborating information (such as genetic risk scores) or high impact incentives, to have the opportunity to potentially improve the correction and/or decisions on which results to reject.

The detailed aims of this statistical investigation are to (i) make a first pass estimation of the apparent dilutions of serum assays from a given sample, using a 'one model fits all assumption' (ii) to consider whether any results should be excluded or corrected because of the dilution problem and (iii) to correct results for assay drift.

Assays are positive-valued and typically have distributions between normal and log-normal. The dilution and calibration effects would be anticipated to have approximately proportional effects across different true values. Therefore, log transformed assay values have been considered in all the correction processes. Assay values may be deemed 'nonreportable' by the lab because the value is below or above the limits of reportability for the assay or for other problems. The following abbreviations are used for the assays:

Short code	Assay
Measured in serum from the serum separation tube sample	
ALB	Albumin
ALP	Alkaline phosphatase
ALT	Alanine aminotransferase
APOA	Apolipoprotein A
APOB	Apolipoprotein B
AST	Aspartate aminotransferase
BILD	Direct bilirubin
BUN	Urea
CA	Calcium
CHOL	Cholesterol
CRE	Creatinine
CRP	C-reactive protein
CYS	Cystatin C
GGT	Gamma glutamyltransferase
GLU	Glucose
HDL	HDL cholesterol
IGF1	IGF-1
LPA	Lipoprotein A
LDLD	LDL direct
OES	Oestradiol
PHOS	Phosphate
RF	Rheumatoid factor
SHBG	SHBG
TBIL	Total bilirubin
TES	Testosterone
TP	Total protein
TRIG	Triglycerides
UA	Urate
VITD	Vitamin D
Measured in red blood cells (plasma depleted EDTA samples)	
HBA1C	Glycated haemoglobin

2.0 Unplanned dilutions investigation

There were three main stages to the process of estimating and partial correction for the unplanned dilutions (i) detection of time periods with differing dilutions (ii) estimation of the apparent dilutions of each assay result and (iii) evaluation of the extent of the problem and application of a correction and exclusions.

2.1 Stage 1: Detection of time periods with differing dilutions

During recruitment, most serum samples were processed using 6 liquid handling machines, each with 8 tips, giving 48 machine-tip combinations. The blood samples collected at the repeat assessment visit were processed using 5 liquid-handling machines. A single tip was used to dispense up to 4 aliquots from a participant sample. A small proportion of samples from each assessment were manually aliquoted. Results were reported as coming from aliquot 1-4, a manually generated aliquot, or this information was missing.

The machine-tips may have operated with undetected faulty seals for several days. Further, upon detecting a faulty seal, the seal may not have been immediately replaced, in part due to a lack of replacement parts. Consequently, it should be possible to identify periods with unexpected dilutions on particular machine-tips by looking at assay results in the order in which they were collected. The precise time of aliquot generation is not available, but in general the aliquots were generated within 24 hours of participant sample collection. Therefore, the date of the participant attending the assessment centre is a reasonable proxy for this information, with a 1 day resolution.

Characterisation of the dilution problem leading to the estimation process

The dilution problem was observed to increase with aliquot number; in addition, assays with a high biological coefficient of variation (CV) are largely unaffected by the dilution problem (in that small percentage dilutions are largely immaterial for assays with a large biological range) and conversely, assays with the lowest biological CVs are most strongly affected.

Only aliquots 1-3 have been included in the dilution estimation process as there were not enough sample results from aliquot 4s for accurate assessment of the dilution problems.

Variable dilutions would not only affect the mean assay values but would also contribute artefactual correlation between the results from different assays on the same sample. The magnitude of this artefactual correlation would in principle depend on the true correlation and the extent to which the two assays were affected by dilution. In practice, the apparent dilutions observed may also be influenced by 'matrix' effects, whereby on dilution by a given percentage with water, different assays do not perform in an entirely *pro rata* manner. Such 'matrix' effects may be particularly likely for assays where the specified diluent (for assaying high values) is not water: this includes 11 of the assays (TES, OES, VITD, IGF1, SHBG, LPA, CRP, RF, APOA, APOB, CYS). For assays with lower and upper limits of reportable ranges that exclude non-trivial numbers of results (e.g., OES and RF were expected to have large numbers of samples with values lower than the reportable range owing to the characteristics of the cohort), the proportions excluded are also likely to be affected (Table A1). For OES, the low analytical sensitivity of the assay meant that samples with naturally very low levels (e.g., those from postmenopausal women) were below the lower reportable range. For RF, the high number of samples with values below the reportable range is also to be expected given its use as a diagnostic marker for rheumatoid arthritis. Non-reportable assay values below or above the reportability limits were included in the change point analysis but not in later stages of analysis and are set to missing in the dataset.

Aliquot	Number of original baseline results below reportable range				
	Lipoprotein A	Oestradiol (females)	Rheumatoid Factor	Testosterone (males)	Vitamin D
Manual	907 (10.5%)	3378 (75.5%)	8247 (90.7%)	7 (0.2%)	76 (0.9%)
Aliquot 1	41822 (10.3%)	157625 (75.2%)	381003 (90.6%)	184 (0.1%)	2302 (0.6%)
Aliquot 2	2921 (10.2%)	12017 (74.5%)	26751 (91.2%)	11 (0.1%)	156 (0.6%)
Aliquot 3	1079 (11.1%)	5311 (74.5%)	9798 (91.5%)	7 (0.2%)	53 (0.5%)

Table A1 Numbers of results below the reportable range for assays with >0.1% of results below the reportable range in baseline samples[1]

[1] No assays had >0.1% of results above their reportable range. For oestradiol, the relatively high proportion of samples with values below the reportable range in women reflects the menopausal status of participants at recruitment (with ~25% being premenopausal; oestradiol results in men were not included here). For rheumatoid factor, the high proportion of samples with values below the reportable range reflects its use as a diagnostic marker for rheumatoid arthritis.

Detection of time periods with differing dilutions

After some exploratory investigations, 7 assays (ALB, CA, CRE, CYS, GLU, PHOS, and TP) from among those most strongly affected by dilution by various criteria, were used in the analysis to detect time periods with different dilutions. A multivariate change point analysis was conducted using the selected 7 assays to identify, for each machine-tip, time points (of sample collection) at which there were jumps in mean assay levels.

Change point analysis

- Performed in 7 selected assays (those most strongly affected by dilution by various criteria) and restricted to samples with complete data on aliquot number and machine-tip;
- Mean values for each machine-tip by collection day were generated;
- Each day was treated as a 7-dimensional observation from a multivariate time series using the ecp algorithm in R 3.5.0 to identify significant change points with the minimum time period set at 3 days. This procedure uses permutation testing (10,000 repeats) to identify significant changes in the distribution and a P-value of 0.01 was used;
- The algorithm was run independently for each machine-tip (48 at recruitment + 40 at repeat visit) x aliquot number (3);
- After running the above analyses, to consolidate the change points for each machine-tip, change points occurring in any aliquot within 3 days of each other were consolidated as a single change point (with the earliest date allocated); in addition periods were merged with adjacent periods where necessary to have a minimum of 30 samples per period.

The change point analysis divided the sample collection timeframe for each machine-tip into periods with distinct assay performance, as shown in Figure A1. This process brought to light a few periods with other anomalies (discussed next) requiring removal of the samples and re-running the change point analysis.

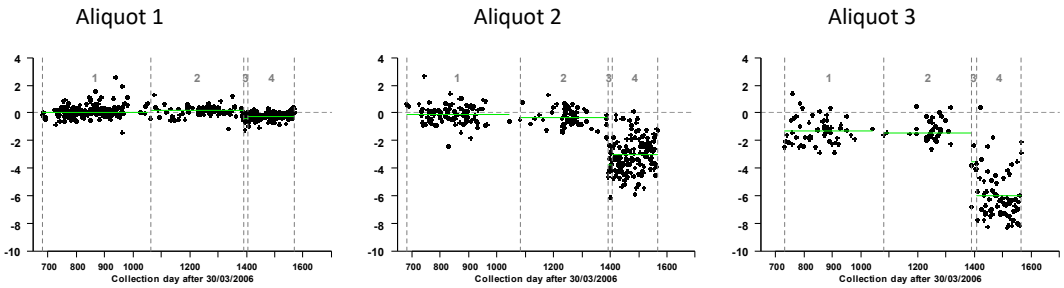

Figure A1. Example of change points identified across the aliquots for a specific machine-tip during one of the worst affected periods

Issues identified: 'Dips' and aliquot misclassifications

Some time periods showed a few other obvious anomalies likely to correspond to severe dilution problems or aliquot number misclassification.

Dips

Visual assessment of some of the assay results identified short periods of time on particular machine-tips when the values were highly variable, even in aliquot 1 (Figure A2). We call such periods 'dips', as the results visually dip very low compared to the overall population. As this could be an indication of a severe dilution problem, results from these samples were excluded from the final estimation and from both the Data Showcase and extended datasets.

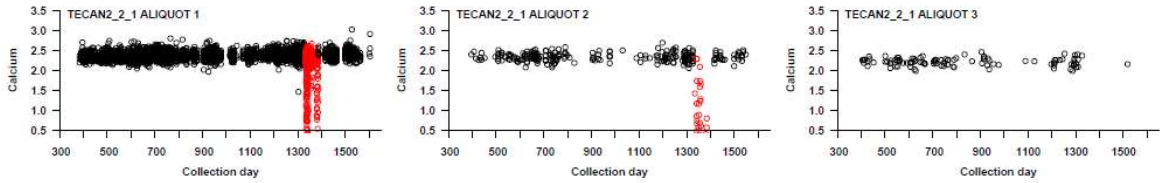

Figure A2. Calcium assay results from a particular machine-tip across the 3 aliquots, showing a large dilution effect in aliquot 1 and 2 for some samples over a small period of time. It appeared that no results were obtained in aliquot 3 in the same period

Possible aliquot number misclassification

Visual inspection of some of the assay results also identified periods on particular machine-tips when assay results appeared to have a bimodal distribution (with the lower valued component being more out of line), suggesting possible aliquot misclassification, meaning that one of the populations of results may have come from a different aliquot to that recorded by the lab (as illustrated in Figure A3). We would expect the proportion of samples from each aliquot to be similar across the tips within a machine, but checking this identified 4 periods with different proportions of results from each aliquot and with a bimodal distribution suggesting they were two distinct populations of results. This was indicative of possible aliquot misclassification.

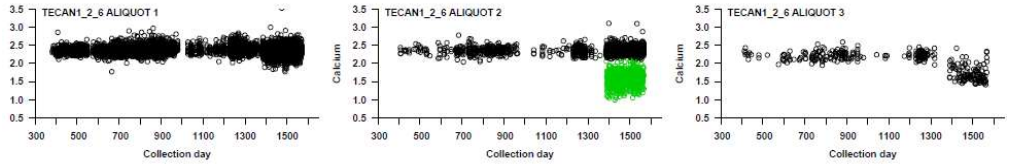

Figure A3. Example of assay results from the same machine across the 3 aliquots, showing a bimodal distribution of results in aliquot 2, suggesting possible aliquot misclassification

Where aliquot misclassification was suspected, the mclust algorithm (implemented in the R package *mclust*) was used to generate estimates for the means and standard deviations (SDs) of a 2 component mixture of Normal distributions. This procedure maximises the Bayesian Information Criteria (BIC) for the mixture, allowing different means and different SDs in the 2 mixtures. The split point, x_s, where there is the same probability that a value belongs to each mixture component is estimated. We reject results from participants with results below x_s, (i.e., estimated to be from the component with the lower mean) suspecting aliquot misclassification.

Results of the change point analysis

The process identified change points in nearly all machine-tips. In the baseline sample, the process initially identified a total of 103, 2, and 1 change points from the analysis of aliquots, 1, 2 and 3, respectively, which were consolidated into 106 change points (after merging change points < 3 days apart and periods with <30 samples). Of the 48 machine-tips, 3 had no change points, 6 a single change point and 39 had two or more change points.

2.2 Stage 2: Estimation of dilution

Dilution estimation exclusions

Details of the exclusion criteria applied for the dilution estimation are provided below. Assay results were excluded if the aliquot had failed by the processes described above, or aliquot 4 was used, or if the analyser provided an unreportable result.

Assays excluded from dilution estimation included those with a high biological CV (LPA, GGT, CRP, RF) plus VITD, where seasonal variation masked any dilution effects related to time of sample acquisition, leaving 24 assays included.

[N.B. The estimates of dilution were only performed in the serum sample. The single assay performed in red blood cells (HbA1c) was not considered in the estimation of dilution as any effect of dilution is unlikely to affect the results, given that HbA1c is presented as a concentration ratio. Additionally, only 1 aliquot was dispensed (from 2 EDTA sample tubes), and hence it's not possible to assess the effect of dilution across different aliquots. As such, the HbA1c assay results do not have a corresponding aliquot number associated with them.

Inclusion criteria

All assays

1. Within reportable range at initial measurement
2. No error reported

Assays on serum samples

3. Not in dip period
4. Not in “aliquot misclassification” period
5. Aliquot number 1-3, manual or missing

Inclusion rules for any dilution factor estimation (for other serum results the original result is retained)

6. In a restricted list of assays (see Table A2)
7. Not more than 15 SDs below mean in manual aliquots

Inclusion rules machine-tip dilution factor estimation (for other serum results the original result is retained)

7. Aliquot number 1-3, manual or missing
8. Has machine-tip information or was a manual sample
9. More than 1 result from machine-tip (a single machine is recorded as being used only once for the resurvey data, so correction was not possible)
10. Not from visits in 2006, a year before the main baseline recruitment, because there were only 4 such samples with data (which is too sparse for this analysis)

Inclusion in sample reduction in concentration estimation

11. With machine-tip dilution factor estimate but from a restricted list of assays (see Table A2)

Inclusion in date correction

12. Results not excluded by 1-5 or 7 above
13. In restricted list of assays listed in point 6 plus HbA1c (for other results the original result is retained)

Main and extended datasets

Main Showcase dataset:

1. Corrected (or uncorrected results as appropriate) results within reportable range before and after any correction.
2. Serum results are rejected for samples with (i) estimated sample reduction in concentration outside (-10% to +10%) or (ii) for assays (ALB, APOA, CA, GLU, HDL, PHOS, SHBG, TP) estimated sample reduction in concentration outside (-1% to +1%)

Extended dataset:

1. Additionally includes corrected results that are outside reportable range after correction (but were reportable originally)
2. No rule applied to reject on the basis of the estimated sample reduction in concentration, but flags indicate if rules applied to a result in the main dataset

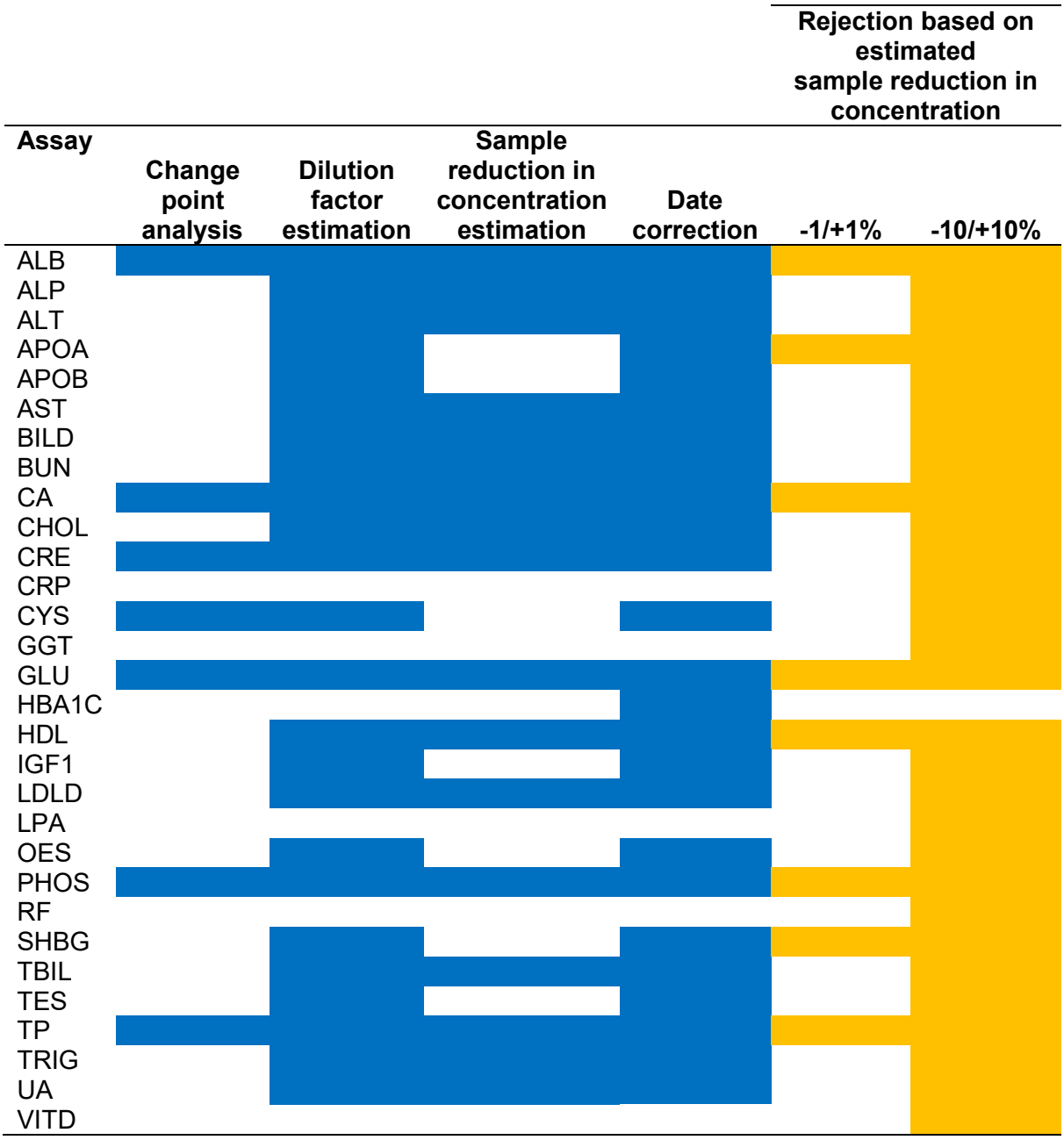

Table A2. Assays included in each analysis stage

Modelling dilution

For each assay result the dilution factor was defined as the multiplicative factor applied to the theoretical true result (i.e. from perfect aliquoting) which would give the observed result, i.e. a dilution factor of 85% means observed=0.85×theoretical true (so 1/0.85≈1.17 means that the system fluid represented a 17% additional volume).

The basic principle of the correction is that, for an observed assay result $Y_{diluted}$

$$Y_{diluted} = \text{dilution\_factor} \times Y_{true} + \text{error}$$

After taking logarithms of the assay results, the multiplicative factor becomes an additive one relative to manual aliquots, and the model becomes

$$\log(Y_{diluted}) = \log(Y_{true}) + \log(\text{dilution\_factor}) + E, \quad E \sim N(0, SD^2)$$

We anticipate that the dilution factor depends on aliquot×machine_tip_period, but in addition, the apparent dilutions observed for different assays will vary somewhat around the actual sample dilution due to substrate 'matrix' effects, whereby different assays do not yield completely pro rata effects for a given dilution with water. Therefore, the model fitted also includes a term for assay×aliquot. As this method is only applicable to variables with equal variance, the terms are divided by the SD of the log assay. TES and OES were included as separate assays for each sex, taking the number of 'assay' categories in the model from 24 to 26. Hence the model fitted was

$$LYS_{ij} = assay_j + assay_j \times aliquot_i + aliquot_i \times machine\_tip\_period_i$$

where:

- LYS_{ij} is the log assay result for assay j (j=1,… 26) for sample i divided by the standard deviation of the log results for that assay
- $assay_j$ is a categorical variable for the assay (j=1,…26) yielding result Y_{ij}
- $aliquot_i$ is a categorical variable for the aliquot number (0-3, where 0 denotes manually aliquoted) for sample i
- machine_tip_period is a categorical variable for the machine tip used for sample i

Separate models were fitted to the baseline and repeat visit results (as the machine-tip-periods were distinct). The parameter estimates were calculated with manual aliquots (~9,000 of the baseline samples and ~6,000 of the repeat assessment samples) as the reference group, but in a further step were referenced to a larger group by the addition of a constant to yield an average dilution factor of 1 in the larger group (defined as participants with manual aliquots plus samples with calcium results from aliquot 1 with an estimated dilution factor of 0.99-1.01: this group contained ~350,000 samples at baseline).

Since LYS_{ij} is log transformed, we can interpret the coefficients as a scaling applied to the original untransformed variables. Therefore, for a given assay result from a given aliquot and machine_time_period, if we estimate β as the $assay_j$×aliquot effect and δ as the aliquot×machine_time_period, then exp(β)×exp(δ) = exp(β+δ) is the estimated apparent dilution factor.

The estimated sample dilution factor was calculated as the exponential of a weighted average with weights $1/SD_j^2$ of the model terms relevant to the dilution in each assay over 17 assays (Table A2), excluding assays with a significant proportion of results below the lower reportable limit and assays where the normal diluent was not water) i.e., for sample i

exp(sum j=1, 17 ($assay_j$×$aliquot_i$ and $aliquot_i$×machine_tip_$period_i$) /SD_j^2 x sum j=1,17 SD_j^2)

We refer to (1-estimated sample dilution factor) x 100 as the estimated percentage reduction in sample concentration.

The model parameters were estimated using PROC GLM in SAS 9.4. 10-fold cross-validation was used to avoid overfitting (i.e. the samples were randomly assigned into 10 groups; for each 10th, the other 90% of the data was used to generate parameter estimates for that 10th of the data).

2.3 Stage 3: Evaluation and decisions on correction and exclusion from Data Showcase

The distribution of estimated sample reductions in concentration by aliquot number and overall is shown in Table A3. Among the baseline samples in the table, 98.5% of aliquot 1 samples had estimated percentage reductions in concentrations within +/- 1%; for aliquot 2 most sample reductions in concentration were in the range 1-3% and for aliquot 3 most were in the range 5-10%. Only 7 samples (all aliquot 3) had estimated reductions in concentration above 10%. Overall, 91.6% of the samples assayed had estimated reductions in concentration within +/-1% or were manually aliquoted.

		Estimated percentage reduction in sample concentration							
Aliquot Number	Number of samples	≤-2%	>-2% to ≤-1%	>-1% to <-1%	≥1% to <3%	≥3% to <5%	≥5% to <10%	≥10%*	Mean
Manual	9086	-	-	100%	-	-	-	-	-
1	418170	0.00%	0.21%	98.54%	1.25%	-	-	-	0.1%
2	29050	-	0.01%	20.55%	77.88%	1.56%	-	-	1.4%
3	10230	0.01%	-	-	0.02%	1.94%	97.97%	0.07%	6.4%
Total	466536	0.00%	0.19%	91.55%	5.97%	0.14%	2.15%	0.00%	0.3%

Table A3: Distribution of estimated percentage reductions in sample concentration by aliquot number (among baseline samples where the same aliquot was used for all assays).

* Results for serum assays with estimated percentage reduction in sample concentration ≥10% are set to missing in the Data Showcase data.

After excluding from consideration assays with a significant proportion of results below their lower limit, the extent of the dilution problem was characterised for the remaining assays included in the dilution estimation by comparing the correlations between assays in manually aliquoted samples with those in samples from a given aliquot number or a given estimated sample dilution range. These differences in correlations were reviewed using heatmap visualisations. Four assays with the lowest biological CVs (CA, TP, PHOS, ALB) showed distorted correlations with each other and to a lesser extent with some other assays. In addition, a further 4 assays (GLU, HDL, APOA, SHBG) also showed substantially distorted correlations with several assays. The distortions tended to increase with aliquot number and with estimated reduction in sample concentration and also with estimated reduction in sample concentration within a given aliquot number, where this was assessable (i.e., where a range of substantial reductions were present within an aliquot number: for aliquot 2, reductions in sample concentration up to 5%, and for aliquot 3, reductions above 3%, Table A3). Therefore, the estimated sample reduction in concentration appeared to add some information on dilution over and above aliquot number, at least in these ranges, and so the full model was adopted as the method of estimating the sample dilution.

However, applying the modelled predictors of the apparent assay result dilution factors (involving about 500 terms) to assay results to correct for dilution, made only a small improvement to the correlation distortions. This failure may be partly because only 3 of the change points identified were in aliquots 2 and 3, which was probably primarily due to low frequency of use of these aliquots. However, it could also be an indication that the dilutions

were varying over a shorter timeframe than could be captured by the present model and data. The majority of the change points were derived in aliquot 1, where the dilutions were small and any improvement from correction would be largely negligible and difficult to evaluate. It is also possible that differences between assays may not have been adequately catered for in this first pass model. Given the limited improvement achieved, the model was rejected as a satisfactory correction for assay results. Instead a further model involving just terms for assay x aliquot number was fitted to correct for differences in apparent dilutions by aliquot number (involving 3 terms per assay). The aliquot-number-corrected result was obtained by dividing the observed result by the estimated apparent dilution factor from this model based on aliquot number.

Results from all assays were excluded from Data Showcase for the 7 samples with estimated reductions in sample concentration outside -10 to 10%, and from the 8 assays mentioned above (that showed the worst distortions in the correlations between assays) for samples with estimated reductions in sample concentration outside -1 to 1%. The result corrected for aliquot number and date of assay is given in Data Showcase and in the extended dataset for the included results.

The various indications for dilution (shift in mean levels, distorted correlations, low biological CV) did not give an entirely consistent pattern across assays and various odd anomalies remain unexplained. Therefore, researchers with expertise in particular assays or interest in pursuing more advanced modelling may be able to take this forward to improve the UK Biobank resource, such as by charactering the reductions in apparent concentration of the excluded results better and potentially reprieving some of these results.

3.0 Date of assay corrections

For each assay, plots of the daily mean results by date-of-assay revealed that many assays showed drifts over time (up or down) and occasionally exhibited step changes. Changes in reagent batches could account for some of the variation but other variation was detectable by the volume of results in comparison to the limitations of standard lab QC procedures.

Some assays (typically with low CV) exhibited large amounts of variability by day of assay. Conversely, some assays with high CVs were seen to have low variability by day of assay in the time series.

Based on this, the seasonal effects on vitamin D and a preliminary run (referred to later in Figure A4), it was decided that the results from all serum assays except for VITD, LPA, GGT, CRP and RF should be corrected for date of assay effects. In addition the date of assay correction process was also applied to HbA1c results, which were from a different type of sample (red blood cells) not investigated for dilution, but which were affected by day-to-day lab variability. The date of assay effects were assumed to be independent of the aliquot dilution problem and the correction was applied after correction for aliquot dilutions.

3.1 Correction for date of assay effects

Least square means (LSMEANs) were generated from a linear model of log assay values (using aliquot-dilution corrected assay results, where available) on date-of-assay as a categorical variable, with adjustment for the interaction of sex with age at survey (as a categorical variable for single year [recruitment ages <40 or >70 were combined into groups for age 40 or 70, respectively]). Age and sex were included as a precaution to allay concerns that differences in the participant characteristics of the samples assayed each day might be contributing importantly to the date-of-assay effects (but the age and sex terms appeared to account for little of the day-to-day variation.) The date-of-assay effect was calculated as the difference of the date-of-assay LSMEAN from the overall mean. Values were corrected by subtracting the corresponding date-of-assay effect (resulting in correction for date of assay but not adjustment for age and sex).

Recruitment and repeat assessment samples were included together in the models for each assay, as these samples were mixed together across assay batches throughout the assay period. Days with <30 observations were grouped with the neighbouring day (forwards or backwards). This was done recursively until all day groups had at least 30 observations.

In a preliminary run, for each assay, the change in within-participant correlation (between recruitment and repeat assessment) with adjustment for date-of-assay was plotted against the assay biological CV (Figure A4). For the majority of assays, adjustment for date-of-assay improved the self-correlation, with the exception of assays with a high biological CV, where adjustment for date-of-assay reduced this self-correlation. In the light of these preliminary observations, assays excluded from the date-of-assay correction were those with a high biological CV (CRP, GGT, LPA and RF) and those with evident seasonal variability (VITD).

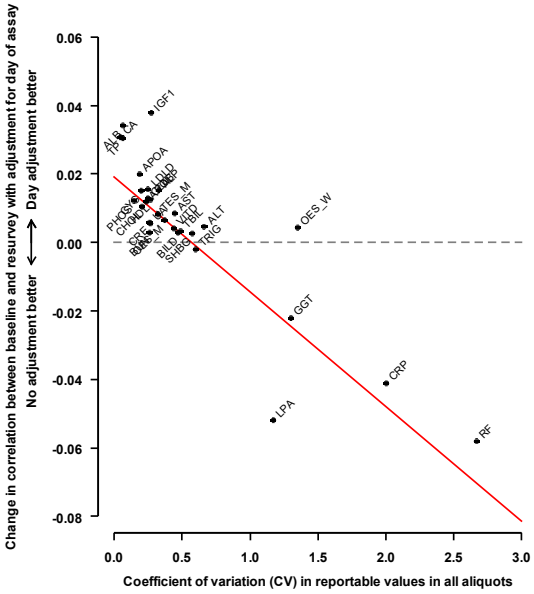

Figure A4. Change in correlation between the baseline and repeat sample for date-of-assay by assay biological CV

Assessment of corrections

The correlations between baseline and resurvey assay values in the same person generally showed a slight improvement with correction for aliquot dilution and date of assay. For example, for a badly affected assay, such as calcium (CA) the Spearman correlation improved from 0.331 uncorrected, to 0.347 with aliquot dilution correction, to 0.377 with date of assay correction, whereas the correlation for total cholesterol only improved from 0.651 to 0.656 to 0.659.

0. Delivered datasets

Notation: <survey> takes values b=baseline or repeat assessment. <Assay> takes the assay abbreviations.

- **Main data showcase**

Variable name	Description
Est_sample_dilutionfactor_<survey>	Estimated sample dilution factor
<assay>_result_<survey>	Assay result
<assay>_date_<survey>	Assay date
<assay>_aliquot_<survey>	Assay aliquot (. [missing],0=manual,1,2,3,4)
Assay status variables	
<assay>_correctionlevel_<survey>	Correction level (missing means no result) • 0: none • 1: date and aliquot correction • 2: date correction only
<assay>_correctionlevelreason_<survey>	Reason for correction level • 0: normal for assay type (i.e. 5 assays are not corrected at all, and HBA1C can only have date of assay correction) • 1: no tip information • 2: no aliquot information • 3: original log assay result is more than 15 SDs below the mean of log results in manual aliquots • 4: resurvey only: insufficient participants with aliquot dispensing machine data for aliquot correction
<assay>_noresult_<survey>	Reason for no result (missing means has result) • 1: no data returned • 2: original value above or below reportable limit. • 3: unrecoverable aliquot problem (dip) • 4: unrecoverable aliquot problem (possible aliquot misclassification) • 5: aliquot 4 used • 7: analyser deemed result not reportable for reason other than above or below reportable range • 8: not reportable because sample dilution factor 0.9-0.99, 1.01-1.1

	• 9: not reportable because sample dilution factor <0.9, >1.1
<assay>_report_<survey>	Reportable status (missing means result not attempted) • 1: reportable at assay and after aliquot correction, if attempted • 2: reportable at assay but not reportable after any corrections (too low) • 3: reportable at assay but not reportable after any corrections (too high) • 4: not reportable at assay (too low) • 5: not reportable at assay (too high)

- **The extended dataset contains the following variables:**

Variable name	Description
Est_sample_dilutionfactor_<survey>	Estimated sample dilution factor
<assay>_result_<survey>	Assay result (after any corrections are applied). As result in main data but also including (i) results reportable originally but not within the reportable limits after correction and (ii) results excluded for <assay>_noresult_<survey> = 8 or 9 in main data.
<assay_result>_in_main_flag_<survey>	Result included in main Showcase data (0/1)
<assay>_oval_<survey>	Original assay result (direct from analyser) for included result
<assay>_date_<survey>	Assay date
<assay>_aliquot_<survey>	Assay aliquot (. [missing], 0=manual,1,2,3,4)
<assay>_dilutionfactor_mtpmodel_<survey>	Assay estimated dilution factor from the model including machine-tip periods
<assay>_dilutionfactor_alomodel_<survey>	Assay estimated dilution factor from the model using aliquots only
<assay>_datefactor_<survey>	Assay estimated date correction factor (computed following the aliquot only model for dilution correction)
Assay status variables	
<assay>_correctionlevel_<survey>	Correction level (missing means no result) • 0: none • 1: date and aliquot correction • 2: date correction only
<assay>_correctionlevelreason_<survey>	Reason for correction level • 0: normal for assay type (i.e. 5 assays are not corrected at all, and HBA1C can only have date of assay correction) • 1: no tip information • 2: no aliquot information

	• 3: original log assay result is more than 15 SDs below the mean of log results in manual aliquots • 4: resurvey only: insufficient participants with aliquot dispensing machine data for aliquot correction
<assay>_noresult_<survey>	Reason for no result in extended dataset (missing means has result) • 1: no data returned • 2: original value above or below reportable limit. • 3: unrecoverable aliquot problem (dip) • 4: unrecoverable aliquot problem (possible aliquot misclassification) • 5: aliquot 4 used • 7: analyser deemed result not reportable for reason other than above or below reportable range
<assay>_report_<survey>	Reportable status (missing means result not attempted) • 1: reportable at assay and within reportable range after aliquot correction, if attempted • 2: reportable at assay but outside reportable range after any corrections (too low) [assay result to be hidden in main release] • 3: reportable at assay but outside reportable range after any corrections (too high) [assay result to be hidden in main release] • 4: not reportable at assay and outside reportable range after corrections (at assay, too low) [assay result to be hidden in main release] • 5: not reportable at assay and outside reportable range after aliquot correction (at assay, too high) [assay result to be hidden in main release]
<assay>_error_<survey>	Flag for whether analyser deemed result not reportable for reason other than above or below reportable range (1 for such results, missing otherwise)
dips_b	Flag (0/1) for baseline sample aliquot during dip period
misclass_b	Flag (0/1) for baseline sample aliquot during possible aliquot misclassification period
machine_tip_<survey>	Machine-tip used for sample aliquots

machine_tip_changepoint_period_<survey>	Machine-tip period used in aliquot correction
noresult_main_code_8_<survey>	Result excluded from main data with <assay>_noresult_<survey>=8
noresult_main_code_9_<survey>	Result excluded from main data with <assay>_noresult_<survey>=9