STEpUP OA replication analysis plan

v1.0, 16/05/2023

Introduction

This document outlines the plan for the analysis of the STEpUP OA Replication dataset.

The Replication proteomic dataset comprises data generated using synovial fluid (SF) samples from tranches 3 & 4. A total of N = 707 samples (from N = 669 patients) were processed on the SomaScan SOMA plex V4.1 platform. This left a total of N = 669 samples from 669 baseline patient samples from 8 cohorts in the Replication data analysis. Baseline samples are defined as a) the earliest biological sample for each participant (this is typically the baseline visit), that b) has a disease group and SIN assigned in the clinical database, and thus included 'baseline' samples from the knee joints of osteoarthritis participants (OA, N=701); N = 595 true baseline samples, N = 20 (visit 3), N = 19 (visit 5), N = 31 (Visit 6), N = 18 (Visit 7), N = 6 (Visit 8) and N = 12 (Visit 9) samples, respectively. Of the baseline OA samples, N=429 were spun and N=235 were unspun. Unlike the discovery dataset, no joint injury cases are included in the replicate dataset. A further N = 38 samples from contralateral knees sampled, of which N=32 were from the same visit and N=6 were from different follow-up visits, were also analysed but are not defined as baseline samples and will be analysed separately. A breakdown of these samples by tranche, centrifuge status and baseline status are summarised in the table below.

	Tranche 3	Tranche 4	Total
Baseline knee OA, spun	419	10	429
Contralateral knee OA, spun	38	0	38
Baseline knee OA, unspun	235	0	235
Contralateral knee OA, unspun	0	0	0

All samples have associated clinical data including core demographics (e.g. participant age, sex), and cohort-level/disease status data (e.g. cohort name, tranche number, disease grouping (i.e sf_iknee_qc_group)). The majority of the samples also have at least one cross-sectional prioritised/harmonsed single measure of knee pain (WOMAC for OA, KOOS for joint injury, or where this is not available a knee-specific VAS/NRS or PainDETECT VAS), and many have a measure of radiographic disease severity available (e.g. ordinal Kellgren-Lawrence (KL) grade, binary radiographic knee OA status, and/or binary advanced

radiographic knee OA status variables). We also have data available on confounders (e.g. BMI and ever smoking history) for most samples.

The Replication dataset for this analysis plan will consist of N = 707 samples with proteomic and clinical data. The technical and clinical variables included in the dataset are the same as in the Discovery Analysis Plan v1.1 data release, and details given in Appendix 2 and 3 (which are mostly the same as the table in the appendix of the Discovery data analysis plan), and descriptive statistics for these variables in the Replication dataset are given in Appendix 1.

Data processing and quality control process

The total of N=707 samples with proteomic data were derived from an initial replication sample set comprising a larger number of N = 728 OA synovial fluid samples. Before proteomic profiling by Somalogic in the rerun, 10 samples were excluded by the Oxford lab due to low insufficient sample volume. Of the remaining N = 718 samples, N = 11 further samples were not processed by SomaLogic due to insufficient sample volume.

SomaScan data processing normalization and quality control will be carried out as specified in the Discovery Analysis Plan v1.1, with some modifications. All the spun sample data underwent optimised standardisation (described in the Discovery Analysis Plan), intracellular protein score adjustment (a modification of the Total Signal Intensity adjustment, described below), batch correction for the category combining the plate and bimodal signal (described in the Discovery Analysis Plan), and finally filtering the samples and proteins with insufficient quality (using a slightly different filtering process to the Discovery Analysis Plan, outlined in Appendix 4 and also included in our QC manuscript). The only change in filtering was the removal of the sample volume filter, which was modified in response to feedback from clinician members of the Data Analysis Group that differences in sample volume often reflected differences in disease severity. Note that, for consistency, all Discovery Analyses have also been rerun using these modified QC procedures.

After starting the discovery analysis, we found that the first principal component, which we originally described using a total signal intensity (TSI) variable, could be well described using an Intracellular Protein Score (IPS) variable, calculated as a weighted sum of log relative concentration (relative fluorescence units), where the weights are given by the Cohen's d of log expression between the spun and unspun paired samples. We have thus replaced the TSI with the IPS as the variable that is adjusted for in the primary adjusted dataset.

Note that the primary replication analysis will only use the spun samples. However, we will also carry out additional analyses to replicate our results in the unspun dataset (to test generalisability), and in the combined spun/unspun data (to maximise power). For analysing the unspun data alone, we will use the same data processing and QC procedure described above. For the combined analysis we will construct a unified dataset composed of spun and unspun samples, which will be processed in the same way but with two modifications: 1) we will adapt the ComBat to batch correction to include spun/unspun status as we well as plate and bimodal signal status, and 2) we will filter proteins that have inconsistent signals in spun and unspun samples. Those proteins will be defined using the 18 spun/unspun paired

samples, and we will remove samples with a nominal (uncorrected) p value > 0.05 in the Pearson correlation test between spun and unspun paired samples.

All analyses will be performed using the standardised, batch-corrected, intracellular protein score adjusted data – these data will be referred to throughout as the 'intracellular protein score adjusted data' (or simply 'adjusted data'), and findings generated using these data will be treated as our primary results. All analyses will also be performed in duplicate using the normalised, batch-corrected data – these data will be referred to throughout as the 'non-intracellular protein score adjusted data' (or simply 'non-adjusted data').

All analyses, unless otherwise stated, will be performed using log-transformed protein abundance data (i.e. log(protein abundance)), and filtered data[1].

Data Analysis Plan

Overview of analysis approach

This replication analysis is designed to replicate and expand on the results of the Discovery Analysis. It is broken down into two sections, Endotype Replication Analysis and Clinical Association Replication Analysis (replicating the results of the Primary Analysis and Secondary Analysis respectively, both described in the Discovery analysis plan).

Not all analyses are strictly replication analyses, as some analyses (e.g. the sex and obesity interactions) are new analyses designed to investigate hypotheses that have arisen since the discovery analysis was written. These new analyses will be run in both discovery and replication datasets.

For all analyses we will (unless otherwise specified):

- only include the baseline sample (earliest sample available, and then using the right knee if bilateral sampling is available)
- when a specific clinical, demographic or QC variable is being analysed, we will remove samples that have missing data for this variable from that specific analysis
- to be carried out only on spun samples, with unspun samples only analysed in the robustness analyses stated in the text
- be carried out on the intracellular protein score adjusted dataset described above, then repeated on the non-adjusted dataset as a robustness test, with similarities and differences between the two analyses reported on.

Throughout the document, all analyses are assumed to be carried out in R[2] unless otherwise specified.

There are four different main datasets adopted in different sections in this analysis plan:

Spun Replication (N=429) -- The primary replication dataset, including all the spun baseline OA samples in the replication release (tranche 3 and 4).

---

[1] Discovery Analysis filters will be applied (these filters are different to those as described in v1.1): n = 6290 proteins included. We will not filter out proteins associated with sample volume.

[2] https://www.r-project.org/

Unspun Replication (N=235) - All the unspun OA samples in the replication release, used to test whether findings generalise to unspun samples (note that there were no unspun samples in the discovery release)
Spun Combined (N=1147) -- All the spun baseline OA samples from both discovery and replication data (all four tranches), used for testing hypotheses in the maximally powered full spun dataset.
Spun+Unspun Combined (N=1385) - All the spun and unspun baseline OA samples across both discovery and replication, after adjustment for centrifugation effects, used to test hypotheses in the largest possible sample size.

In parts of the replication analysis disease-free (DF) controls (N=37), taken from the discovery dataset, are used which are a mixture of spun (N=6) and unspun (N=31) samples from OA-free patients. Other analyses also use the contralateral knee samples (N=32), which are spun samples included in the replication dataset, generated from SF taken from the contralateral knee at the same visit as a baseline ipsilateral knee sample.

Analysis 1: Endotype Replication Analysis

The main purpose of the Endotype Replication Analysis is to verify the robustness and generalizability of the findings discovered when carrying out the Primary Analysis described in the Discovery Analysis Plan v1.1.

Key findings from the Discovery Analysis:

1. After the data were adjusted for intracellular protein score (IPS), we did not find significant clustering within OA patients.
2. We found two significant clusters in OA using the non-IPS-adjusted dataset, with a large number of proteins significantly higher in one cluster than the other (but no proteins significantly lower in this cluster).
3. The two clusters differed significantly on their IPS, and the IPS formed a continuum between the two clusters without a clear cut division between the two
4. We found a number of enriched pathways in differentially abundant proteins between the two clusters, which was true even if IPS was included as a covariate in the differential abundance test. It was not clear whether this was due to residual confounding.
5. We did not find any clinical features correlated with the non-IPS-adjusted clusters

Key Questions to Address in the Replication Analysis:

1. Do we replicate the absence of significant clustering in the independent *Spun Replication* samples, or in the larger, maximally powered *Spun Combined* or *Spun+Unspun Combined* datasets of discovery and replication samples?
2. Do we replicate the two clusters in the non-IPS-adjusted data in the *Spun Replication* dataset? I.e. is there significant clustering, does it have the same characteristic proteins and pathways as in the discovery OA set, and does it also reflect a continuum of IPS value between the two clusters?
3. If clustering does exist in the *Spun Replication* dataset, is it also uncorrelated with clinical features?

4. Do the clustering results for IPS-adjusted and IPS-unadjusted discovery samples generalise to the *Unspun Replication* samples processed in the same way?
5. Is the clustering structure maintained when we subset the patients to different groups based on the clinical feature of advanced radiographic knee OA status, using the *Spun Combined* data to maximise power?

The detailed analysis approaches and required outputs are listed as follows, structured by five sub-analyses to address the major questions listed above.

Sub-analysis 1.1: Endotype Detection in the Intracellular Protein Score Adjusted Data

Questions:

- Does the finding of no significant clustering after IPS adjustment replicate in the independent *Spun Replication* data?
- Does the finding of no significant clustering after IPS adjustment replicate in the maximally powered *Spun Combined* and S*pun/Unspun Combined* datasets?

Results:

- Principal components for each sample
- UMAP coordinates for each sample
- The value of f(K) statistic for each possible number of clusters

Plots:

- Plots of the f(K) statistic against cluster numbers.
- PCA plot coloured by clusters and by intracellular protein score
- UMAP plot coloured by clusters and by intracellular protein score

Method:

To replicate findings from discovery analysis, we will calculate the f(K) statistics and cluster data using k-means clustering on the reduced PCA space.

For the *Spun Replication* data, we will use two different approaches to reduce the dimensions for clustering. One approach will be to perform PCA directly on the protein expression profile of the *Spun Replication* data. The other approach will be to project the *Spun Replication* samples onto the PCA space derived from the discovery analysis. We will test both approaches. To replicate finding in the *Spun Combined* and *Spun+Unspun Combined* data, the dimensionality reduction will only use the direct clustering on the full dataset under analysis. The top PCs will be defined by those with cumulative variation explained accounting for 80% of the total variation.

Combining the two variables (dataset used and PCA reduction approach used), we will calculate f(K) statistic and perform clustering in four different PCA datasets:

- Spun Replication – clustering of PCs generated directly on the *Spun Replication* dataset
- Coordinate Replication – clustering of coordinates of projection of *Spun Replication* dataset to PCA space of discovery analysis

- Spun Combined analysis – clustering of PCs generated directly from the *Spun Combined* dataset
- Spun+unspun Combined analysis – clustering of PCs generated directly from the *Spun+Unspun Combined* dataset

We will use the same criterion to decide whether there exists significant clustering as we used in the discovery analysis, i.e. the data is not significantly clustered if there is no cluster number K > 1 resulting in f(K) < 0.85. If the data are significantly clustered, we will pick the optimal cluster number by majority vote across different clustering metrics (as implemented in the R package *NbClust*)

Sub-analysis 1.2: Endotype Detection before Intracellular Protein Score Adjustment

Questions:

- Do we replicate the finding of two significant clusters in the non-IPS-adjusted *Spun Replication* dataset?
- Are any clusters detected characterised by a continuum of intracellular protein scores between the two in the *Spun Replication* dataset?
- Are the same set of proteins significantly differentially expressed across endotypes in the *Spun Replication* data compared to discovery data?
- Do the same bioinformatic features (significantly differentially regulated pathways, cell types) characterise the clusters in the *Spun Replication* dataset as in the discovery analysis?
- Are the endotypes in the *Spun Replication* dataset uncorrelated with clinical features, like in the discovery data?
- Is the same set of technical confounders associated with the endotypes in the *Spun Replication* dataset as in the discovery analysis?

Results:

- Cohen’s d and p value of t-test (if two clusters are detected) or Cohen's f and p value of one way ANOVA (if more than two clusters are detected) to determine the significant difference between/across the means of intracellular protein score
- A table of differential expression test statistics (p-value and odds ratio) for each protein per cluster, with separate results conditioned and not conditioned on intracellular protein score
- Predicted endotype for samples of replication data by classifying to the nearest cluster centroid derived from the discovery analysis
- Enrichment tables showing lists of pathways, cell types, subcellular locations enriched in each endotype and their corresponding p-values
- A table with significant upstream regulators for each endotype and their corresponding p-values
- Lists showing the common and different significant bioinformatic characteristics between the findings based on discovery data and replication data
- A table showing associations between each proteomic cluster and each clinical feature: p values on each endotype for each clinical feature
- Lists of common/different significant clinical features associated with endotypes between discovery data and replication data

- A table showing associations between each proteomic cluster and each technical confounder: p values on each endotype for each confounder
- Lists of common/different technical confounders associated with endotypes between discovery data and replication data

Plots:

- Violin plots of intracellular protein score distribution across endotypes
- UMAP visualisation coloured by intracellular protein score and shaped by endotypes
- Point graphs showing comparisons of protein differential expression strength across endotypes (p values/odds ratios) between replication data and discovery data
- Venn diagrams showing the amount of common/different significant bioinformatic characteristics/clinical features/technical confounders derived from discovery data and replication data.

Methods:

We will carry out k-means clustering and use the f(k) statistic to assign significance as described in the Spun Replication part of the previous section, applied to the non-IPS-adjusted dataset. To test whether the clustering is consistent between the Discovery and Replication analyses, we will also generate a projected PCA space as described in the Coordinate Replication part of the previous section, and then assign Replication samples to Discovery clusters using nearest centroid classification. We will assess the similarity of the clustering structures by the adjusted rand index based on the sample membership of the clusters. Adjusted rand index > 0.9 will be taken as good evidence that the endotype structure generalises to a broader sample population.

To investigate whether the clusters are featured by the continuum of intracellular protein score, we will visualise the clusters on UMAP coloured by intracellular protein score.

We will carry out the following analyses for the *Spun Replication* analysis (using both the Replication and Coordinate replication clusters) using the same approaches as the Discovery Analysis Plan v1.1:

- Approaches for protein differential expression analysis (Sub-analysis 1.2 in the discovery analysis plan)
- Approaches for bioinformatic characterisation of endotypes (Sub-analysis 1.3 in the discovery analysis plan)
- Approaches for association tests between endotypes and clinical features (Sub-analysis 1.4 in the discovery analysis plan)
- Approaches for association tests between endotypes and technical confounders(Sub-analysis 1.5 in the discovery analysis plan)

We will compare the clusters and cell types by plotting effect sizes (log odds ratio and NES for proteins and pathways respectively) between the two analyses, as well as measuring the amount of overlap between significant (Benjamini-Hochberg adjusted $p < 0.05$) associations in the discovery and replication.

To compare the set of differentially expressed proteins between endotypes in the replication analysis to those in the discovery analysis, we will make the Venn diagram to track the

common and different protein signatures. We will also make a dot plot of p values per protein derived from the differential expression regression model for visualisation.

To compare the significant clinical features associated with the endotypes in the replication analysis and discovery analysis, we will make the Venn diagram to show the common and different significant clinical features.

Sub-analysis 1.3: Impact of Sample centrifugation on the OA Endotype Detection based on Proteomic Profile

Questions:

- Do we replicate the finding of no significant clustering after IPS adjustment in the *Unspun Replication* dataset?
- Do we replicate the finding of two significant clusters before IPS adjustment in the *Unspun Replication* data?
- Are any clusters detected characterised by a continuum of intracellular protein scores between the two on the *Unspun Replication* dataset?
- Is the intracellular protein score confounds the same set of proteins which change their significance status of differential expression across clusters in the *UnspunSet data* , compared to discovery analysis?
- Do the same bioinformatic features (significantly differentially regulated pathways, cell types) feature the clusters in *Unspun Replication* data as in the discovery analysis?
- Are the endotypes in the *Unspun Replication* data uncorrelated with clinical features, like in the discovery data?
- Is the same set of technical confounders associated with the endotypes in the *Unspun Replication* data as in the discovery analysis?

Results:

- Table of coordinates of principal components
- UMAP coordinates
- The value of f(K) statistic for each possible number of clusters
- Enrichment tables showing lists of pathways, cell types, subcellular locations enriched in each endotype and their corresponding p-values
- Tables showing associations between each proteomic cluster and clinical feature/technical confounder: p values on each endotype for each confounder.

Plots:

- UMAP visualisation coloured by intracellular protein score and shaped by endotype
- Point graphs showing comparisons of protein differential expression strength across endotypes (p values/odds ratios) between *UnspunSet* data and discovery data
- Bubble charts of enriched gene sets (pathways, cell/tissue types, etc) comparing results from *UnspunSet* data and discovery data

Methods:

To test whether the “Key findings from the Discovery Analysis” can be generalized to unspun samples, we will perform the series of analysis – unsupervised clustering, identify characteristic proteins of clusters, bioinformatic characterization of clusters, clinical

characteristics of endotypes, correlation of technical confounders with endotypes on the *Unspun Replication* set only.

We will consider that centrifugation status has no impact on OA endotype detection, compared to discovery analysis, if
– there is no endotype detected after IPS adjustment
– two endotypes detected before IPS adjustment
– the same set of proteins significantly differentially expressed between the two endotypes
– the same set of significantly enriched pathways and cell types
– no clinical features characterise the two endotypes

Otherwise we will consider that spinning has an impact on OA endotype detection, and that results from combined spun/unspun datasets should be treated with caution.

Sub-analysis 1.4: Clustering Structure of Subgroups of Patients Stratified by Advanced Radiographic Knee OA Status

Questions:

- Is there significant clustering within subgroups of patients based on disease stage or severity?
    - Specifically, is there clustering within early OA patients (defined as KL grades 0 or 1), established but non-advanced radiographic OA patients (KL grade of 2) or advanced radiographic OA patients (KL grade 3 or 4)?
- Is the clustering structure within these disease stages similar or different to that of OA patients as a whole?
- What are the protein signatures for the clusters (if any) within patients at these different OA disease stages?
- Which pathways/cell types are significantly enriched for the clusters (if any) within patients at these different OA disease stages?
- Which upstream transcription factors are significantly (up/down) regulated for clustering (if present) within patients at these different OA disease stages?
- Any clinical features are significantly associated with the clusters (if any) within patients at these different OA disease stages?
- Any technical confounders are significantly associated with the clusters (if any) within patients at these different OA disease stages?

Results:

- Table of coordinates of principal components
- UMAP coordinates
- The value of f(K) statistic for each possible number of clusters
- Enrichment tables showing lists of pathways, cell types, subcellular locations enriched in each endotype and their corresponding p-values
- Tables showing associations between each proteomic cluster and clinical feature/technical confounder: p values on each endotype for each confounder.

Plots:

- UMAP visualisation coloured by intracellular protein score and shaped by endotype

- Point graphs showing comparisons of protein differential expression strength across endotypes (p values/odds ratios) between data and Spun Combined data
- Bubble charts of enriched gene sets (pathways, cell/tissue types, etc) comparing results to discovery data

Methods:
We will construct three datasets stratifying patients by their radiographic knee OA status from the *Spun Combined* dataset - “early OA”, “established but non-advanced radiographic OA” and “advanced radiographic OA”. Early OA will be defined as having 0 for the radiographic_knee_oa flag, established non-advanced OA will have 1 for the radiographic_knee_oa knee OA flag and 0 for the kl_grade_advanced flag, and advanced OA will have 1 for the kl_grade_advanced flag. Patients with missing or not known for either of these flags will be excluded.

The total number of baseline spun samples in these three groups are:

OA stage	**Dcovery**	**Replication**	**Total**
Early (KL = 0-1)	54	76	130
Established non-severe (KL = 2)	68	66	134
Severe (KL = 3-4)	580	258	838

The same approaches will be taken to define significant clustering structure, cluster samples, perform bioinformatic enrichment tests, and test associations with clinical features and technical confounders, compare the similarity of clustering structure, as described in sub-analysis1.1 and sub-analysis1.2.

Analysis 2: Clinical Association Replication Analysis

The main purpose of the Clinical Association Replication Analysis is to verify the robustness and generalizability of the findings discovered when carrying out the Secondary Analysis described in the Discovery Analysis Plan v1.1.

Key findings from the Discovery Analysis:

1) Using the intracellular protein score adjusted data (adjusting for age and sex), we observed protein differential abundance between OA and disease-free (DF) control (largely unspun) samples. Specifically, N = 2088 proteins were differentially abundant in OA (upregulated: n = 975, downregulated: 1113). Some of the key associated proteins included: fibronectin, VEGF-β, sTREM-1 were upregulated, and sFRP-3 was downregulated in OA.

2) Few proteins (N = 5) were significantly associated with WOMAC pain subscores in OA samples (e.g. NOE2, TBC25, GLYG2, NAR3).

3) Over 100 proteins (N = 191) were associated with KOOS pain subscores in injury samples (e.g. multiple members of the VEGF family were upregulated with low KOOS pain scores: VEGF, L-VEGF165, VEGF121).

4) No proteins were associated with the PASS score in OA and injury samples, respectively (in the primary model).

5) Of those proteins associated in OA, many (n = 947, 45.4%) proteins were also associated with advanced radiographic knee OA status. Some key proteins that were associated (i.e. top 20 most strongly associated at p_adj <0.05) in both OA and with advanced radiographic disease status included: sFRP-3, Fibronectin, MMP-1, Tenascin, PENK, TSG-6 etc.

6) For most clinical features, when we further adjusted for the effects of cohort, either i) all signals for protein regulation were lost (e.g. WOMAC pain score, advanced RKOA status, OA vs DF-controls) or ii) the count of significantly associated proteins dramatically changed. For example, for KOOS pain subscore, N = 191 proteins were associated (N = 49 associated with high KOOS scores, N = 142 associated with low KOOS scores) in the primary model (adjusted for age and sex only); however, further adjusting for cohort increased the number of significantly associated proteins to N = 433 (105 vs. 328).

Key objectives of the Replication Analysis:

- All analyses as listed in subanalysis 2.1 of the Discovery Analysis Plan (v1.1) will be run in the *Spun Replication* dataset (N = 429). Additional analyses will be carried out, as described below, on the *Unspun Replication* dataset (N=235) to test generalisability of the results. Further analyses will be carried out on the DF controls and the contralateral knee samples. Once all analyses have been completed, a further analysis will be carried out on the *Spun Combined* dataset (N=1148) to produce a maximally powered but unreplicated (and thus provisional) analysis.
- We will compare; i) the counts of proteins associated (at p_adj[3] ≤0.05) with each respective clinical feature per disease group (where appropriate), ii) the ratio of upregulated (log odds ratio ≥0) vs. downregulated (log odds ratio <0) proteins associated (at p_adj ≤0.05) with clinical features, and iii) the pathways identified in subanalysis 2.2 across Discovery, Replication, and Discovery + Replication datasets.

For a protein to be defined as being 'replicated' for a given outcome:

A protein's association with a given phenotype will be considered to be successfully replicated if it has a Benjamini-Hochberg (BH) adjusted p-value < 0.05 in both the Discovery and Replication datasets, with effects in the same direction. We will run BH correction across

[3] p_adj defined as Benjamini-Hochberg adjusted p-values (generated from regression modeling). Benjamini. Y, and Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society. Series B (Methodological) 57, no. 1 (1995): 289–300.

all proteins in the Replication dataset; replicability will be defined if the given protein meets p_adjust ≤ 0.05 in both Discovery and Replication datasets.

Effects that are present only in the discovery or replication will be considered unreplicated, and thus provisional, results pending future replication. Note that these are not necessarily false, particularly for analyses (e.g. of ordinal KL grade) for which the replication dataset has higher power.

We will assess replication by:

1) Counting the number of proteins that are significantly (p_adj ≤0.05) associated in both Discovery and Replication datasets

2) Counting the number of proteins that have effects (i.e. log odds ratios <0 or >0) in the same and in opposite directions, stratified by adjusted p-value. If the results generalised well to the replication, significantly more than 50% of proteins should go in the same direction (determined by a binomial test), and for significant proteins almost all (>95% for adjusted p-value < 0.05) should go in the same direction.

3) Assessing how well the beta estimates are correlated for a given clinical outcome, per protein

4) Generating Venn diagrams to illustrate the overlap in proteins/pathways for each outcome

5) Of those proteins differentially abundant in OA, assess how many are also associated with radiographic measures of disease (e.g. advanced radiographic knee OA status).

6) To ensure that our results are not driven by confounding by cohort, we will consider a result to be significant after accounting for cohort confounding if it is either or both:

    a) Significant in discovery (not conditional on cohort) AND significant in replication (conditional on cohort)
    b) Significant in discovery (not conditional on cohort) AND significant in replication (not conditional on cohort) AND significant in the Discovery + Replication analysis (conditional on cohort)

To standardize the number of proteins (SOMAmers) investigated across both working datasets (i.e. adjusted and non-adjusted), the same protein (i.e CellularCompositionScoreReg_filter == “PASS”) and sample filters (i.e. CellularCompositionScoreReg_filter == “PASS”) will be applied to both datasets (i.e. spun and unspun datasets)..

Sub-analysis 2.1: *Finding Knee SF Correlates of Clinical Features using regression modeling*

The secondary analysis is designed to build a reference set of proteins (subanalysis 2.1) and pathways (subanalysis 2.2) that are associated with clinical features of OA in knee synovial fluid (SF). The replication dataset comprises OA samples only.

Question(s):

- Do associations between specific proteins and clinical features generalise to an independent replication dataset? Specifically, of those proteins that are associated with a given clinical feature in the Discovery analysis, are they: i) significant, ii) in the same direction (based on log odds ratio) and iii) of the same magnitude (size of log odds ratio & adjusted p.value) in the Replication analysis?
- As was observed in the Discovery analysis, are any observed associations in the Replication analysis driven by effects of confounding by cohort, and can we determine which signals remain robust after conditioning on cohort? Do these align with the findings of the Discovery analysis?
- Are differences, if any, in proteome profiles across disease groups (e.g. OA vs. DF-controls) driven by sample centrifugation status (spun or unspun)?
- Do the correlations between protein concentration and cross-sectional clinical features vary between men and women, and between obese and non-obese patients?
- Are the SF proteins associated with cross-sectional structural severity and knee-specific pain across individuals also associated with differences in structural severity and knee-specific pain in different knees of the same individual?

Clinical outcomes include[4]:

1. **OA-related outcomes (i.e. to be explored in OA[5] samples only)**

- Continuous WOMAC pain subscore (0-100, 100 = worse possible pain)
- Ordinal Kellgren-Lawrence (KL)[6] Grade (0 = KL0, 1 = KL1 etc.) (reference group: KL grade 0)
- Binary Radiographic Knee OA Status (0 = KL<2, 1 = KL ≥2) (reference group: KL grades <2)
- Binary Advanced Radiographic Knee OA Status (0 = Kl grades 0-2, 1 = KL grades 3-4)(reference group: KL grades 0-2).

2. **Disease Grouping:**

- Binary disease indicator for OA vs. disease-free control status (reference group: DF-controls)

** In specific analyses, i.e. OA vs DF-controls, most DF-control samples were ‘unspun’. Therefore, any differences in proteomes between disease groups could be a result of technical variation. As the Replication analysis comprises a larger number of unspun OA samples (>200), we will compare the proteomes of OA and DF-controls samples in two sensitivity analyses;

i) OA vs DF-controls using only spun samples from both disease indicator groups

[4] The Replication dataset comprises only OA samples.

[5] Disease grouping is based on the variable ‘sf_iknee_qc_group’.

[6] Kellgren J & Lawrence J. Radiological Assessment of Osteo-Arthrosis. Ann Rheum Dis. 1957;16(4):494-502. doi:10.1136/ard.16.4.494.

ii) OA vs DF-controls using only unspun samples from both disease indicator groups

The primary comparison of OA & DF-control proteomes will include both spun & unspun samples, using the spin-status-corrected dataset. The analysis of OA vs. disease-free controls should be interpreted with caution given the small sample size of the disease-free control group, and the origin of the SF in such cases (e.g. from the contralateral knees of unilateral OA patients in some cases).

3. Other outcomes:

- Harmonized pain category (0 = acceptable levels of pain, 1 = unacceptable levels of pain) (reference group: 'acceptable' pain) – this is based on the patient acceptable symptom state (PASS)[7]
- Phenotype grouping (i.e. no pain & no radiographic knee OA, radiographic knee OA only, pain only, pain with radiographic knee OA).

It was decided that investigation of protein abundance against phenotype grouping would not be conducted given the small sample sizes of the non-radiographic/non-painful/symptomatic radiographic disease phenotype groups.

In addition to the clinical outcomes explored in the Discovery Analysis Plan (v1.1), we will include additional investigations of other pain PROMS in the Replication analysis. This will include exploring the relationship between protein abundance and knee-specific NRS and painDETECT where these are available, respectively. Knee-specific NRS was not prioritized as the primary pain PROM for generating the PASS for either OA or knee injury disease groups, so the findings generated from exploring the relationship between protein abundance and WOMAC pain subscore (for OA), and KOOS pain subscore (for injury), respectively, will continue be treated as the primary pain PROM findings.

4. Interaction tests with sex and BMI

In order to test whether the correlation between protein levels and clinical variables are modified by other covariates, we will carry out interaction tests for the OA-related outcomes listed above. The same regression model will be used as described above, but with the addition of a protein*covariate term, and each protein will be tested for a significant interaction. This will be carried out twice, once where the covariate is sex, and once where the covariate is a binary obesity flag (BMI $\geq$ 30), using only individuals with BMI data present. These analyses will be carried out independently in the discovery and replication spun datasets, corrected for multiple testing using the Benjamini-Hochberg procedure, and results will be considered to be replicated if they are significant (p_adjusted < 0.05 in both discovery and replication).

5. Correlation between proteins and OA phenotypes in ipsilateral and contralateral knees

Cross-sectional correlations between SF proteins and knee phenotypes (structural severity, pain) could be driven by systemic (i.e. individual-level) or localised (i.e. knee-level) biological

[7] Georgopoulos V *et al*. Harmonising knee pain patient-reported outcomes: a systematic literature review and meta-analysis of Patient Acceptable Symptom State (PASS) and individual participant data (IPD). Osteoarthritis Cartilage. 2023 Jan;31(1):83-95.

effects. In order to test which is more important, we will use the N=32 contralateral knee samples in the data. For each of these 32 individuals, we will generate outcomes based on the difference in phenotype across the knees, specifically:

a) Difference in advanced OA status (+1 for advanced radiographic OA in ipsilateral knee and non-advanced radiographic OA in contralateral knee, 0 for same status in both knees, -1 for non-advanced radiographic OA in ipsilateral knee and advanced radiographic OA in contralateral knee)
b) Difference in radiographic OA (calculated in the same fashion)
c) Difference in KL grade (KL grade in ipsilateral knee minus KL grade in contralateral knee
d) Difference in WOMAC pain (pain in ipsilateral knee minus pain in contralateral knee)

In each case, we will use as the predictor the difference in log protein expression, and results will be tested in a linear model with no additional covariates (as this analysis is already controlled for confounding by using matched knees from the same individuals).

As power is limited by the sample size, we will only examine 10 proteins per outcome, chosen as the 10 most significant proteins from the cross-sectional analysis. We will use BH correction to test for significance (adjusted $p < 0.05$), and will also plot the effect size from the cross-sectional analysis against the effect size from the ipsilateral/contralateral model to see if they have consistent effects.

Data:

- All analyses listed in subanalysis 2.1 will be performed twofold (with the exception of the OA vs. DF-controls analysis); i) with spun samples only (*Spun Replication*) and iii) unspun samples only (*Unspun Replication)*. We will treat analyses using 'spun' samples only as our primary results.

Early Investigation:

Early investigations comparing log-fold change values generated from linear regression modeling (using *limma* package) against log-odds ratios generated using logistic regression modeling showed strong, linear agreement in generated p-values, but poor agreement in regression estimates – see plot below. This was important when ordering proteins based on both p-values and regression estimates as this yielded different lists (e.g. 'top 10' proteins) when using either log-odds or log-fold change values. To improve the agreement with log-fold change values, it was decided that all log odds ratios would be normalized to per standard deviation change in protein abundance. To achieve this, all protein abundance values will be normalized per standard deviation change in protein abundance using the following function; *function(x) exp(log(x)/sd(log(x)).*

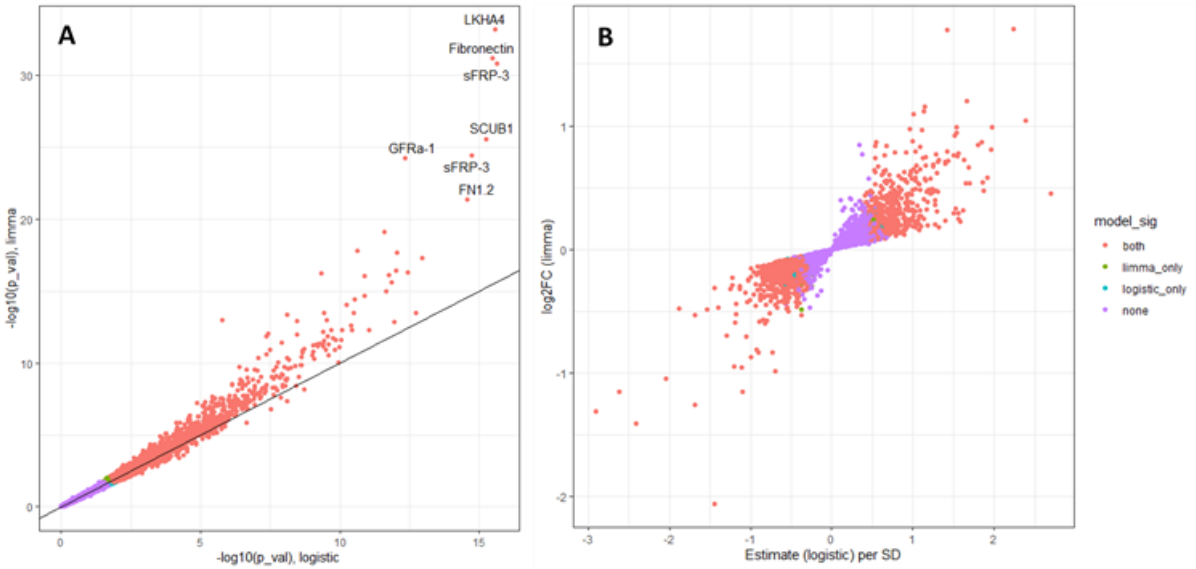


Figure 2: Comparison of (A) p-values from logistic and linear (i.e. limma) models and (B) log-fold change values against log-odds ratios per standard deviation change in protein abundance, generated from linear and logistic regression modeling for the following model: protein abundance against disease grouping (OA vs. DF-controls, adjusted for age & sex). Proteins that were identified as being statistically significant in both models are shown in red.

Methods:

- In each of the respective regression models, we will test log protein abundance as the predictor against each of the clinical outcomes.
- When testing for differences in protein abundance between disease groups (i.e. OA vs. disease-free-controls), as defined using the disease grouping variable ('sf_iknee_qc_group'), we will use logistic regression. When testing for protein abundance differences across the continuous measures including WOMAC pain subscore and KOOS pain subscores we will use linear regression, and for ordinal measures (KL grade), we will carry out ordinal regression.
- Specifically, for linear regression models, we will use quadratic transformations where there is evidence of non-linearity (e.g. lm(womac_pain ~ protein_abundance + protein_abundance, data=data)).
- The primary model will include confounder adjustment for: participant age (at time of sampling) and sex only.
- Further testing will be performed to examine if any of the observed associations are driven by known clinical or technical confounders. Testing will be performed by including different lists of confounders in the regression models;

  - ➢ Primary model: participant age & sex
  - ➢ Robustness test (1): participant age, sex, BMI and smoking history (smoking_history)
  - ➢ Robustness test (2): participant age, sex & intracellular protein score[8]
  - ➢ Robustness test (3): participant age, sex & cohort
  - ➢ Robustness test (4): participant age, sex & log-transformed haemoglobin[9]

[8] Defined as per our intracellular protein score equation.

[9] Log-transformed haemoglobin: calculated as log(seq.4915.64).

➢ Robustness test (5): participant age, sex & centrifugation status

- When testing for differences in log-transformed protein abundance between disease groups, we will carry out a secondary analysis using a random intercept term for cohort in a mixed model using the R package lme4[10].
- Further, we will assess the effects of technical confounders (i.e. plate number, plate position, plate run date, tranche, sample freeze thaw cycles, processing batch, processing date, sample age and sample volume).
- In the Discovery Analysis, we observed in many cases (e.g. differences in injury and DF-control proteomes) strong associations between participant age and the given clinical outcome. Due to the structure of the data and nature of the cohorts included, participant age in such cases almost exclusively explained belonging to a given disease group beyond differences in protein abundance – this was due to little overlap in participant age across e.g. injury and control disease groups, which in turn resulted in fitting errors in logistic regression. Therefore, when model fitting fails due to the occurrence of probabilities of 0 or 1 in the primary logistic model (adjusting for sex & age), we will investigate the sole affect of participant age (by removing as a confounder) and determine whether inclusion of this confounder is appropriate. If we observe either extreme binarisation of adjusted p-values or p-values close to 1.00 or 0, we will remove participant age from the primary model. In addition, in the Discovery analysis most signals for protein regulation were lost after adjusting for cohort. To test the robustness of our findings to potential cohort effects, we will additionally compare: i) agreement in regression estimates and adjusted p-values, respectively, generated from the Discovery (of models adjusting for age & sex only) and Replication (of models adjusting for age, sex and cohort) datasets. In addition, we will perform an analysis pooling all samples from the Discovery and Replication datasets adjusting for age, sex and cohort.

Results:

- Regression coefficients (normalized per standard deviation change in protein abundance), p-values and standard errors for each protein against each clinical feature will be calculated using the appropriate regression model (i.e. linear, logistic or ordinal logistic regression).
- In addition, for each outcome examined, adjusted P-values will be calculated using Benjamini-Hochberg[11] multiple testing correction.
- Protein lists will be generated for each of the given clinical features. These protein lists will not be filtered to only include proteins that are statistically significantly associated (at an adjusted p-value of 0.05) with the given feature.

Plots:

- For each respective clinical outcome, volcano plots of log odds-ratios per standard deviation change in protein abundance against $-\log_{10}$(adjusted p-values) for each protein will be generated. Where appropriate, the top 20 most statistically significant proteins ($p\_adj \leq 0.05$) will be labelled – these plots will be included in the Replication Results release.

[10] https://cran.r-project.org/web/packages/lme4/lme4.pdf

[11] Benjamini, Yoav, and Yosef Hochberg. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." Journal of the Royal Statistical Society. Series B (Methodological), vol. 57, no. 1, 1995, pp. 289–300.

- For each clinical outcome, plots evaluating agreement in proteins that do and do not reach levels of statistical significance (at adjusted p-values of ≤ 0.05) when using either intracellular protein score adjusted or non-IPS-adjusted datasets will be generated – these plots will be used for diagnostic purposes and will be included in the Replication results release.

Sub-analysis 2.2: Bioinformatic characterisation of clinical features

Questions:

- Does the pathway result for each clinical feature generalize to the Replication dataset?
- Do we observe the same pathways associated with clinical features across datasets (i.e …. vs. non-IPS adjusted)?
- Do we observe the same pathways associated with clinical features in spun and unspun samples?

Methods:

The overall approach for subanalysis 2.2 is to test for differences in gene enrichment between disease groups (i.e. OA vs disease-free controls) and across levels of disease severity (e.g. continuous pain measures, ordinal KL grade, binary radiographic knee OA status etc) within OA using log-odds ratios and p-values generated from regression modelling in sub-analysis 2.1. Protein set enrichment testing will be performed using the *fgsea*[12] package in R. Specifically, proteins will be ordered in a ranked list by a 'rank metric' calculated as;

rank metric = -log(p-values) * sign(log odds ratio per SD)

The ranked protein list will then be compared to a gene set (i.e. list of genes known to be associated with a biological process, gene ontology, molecular function or pathway). The 'rank metric' will be used to calculate the normalized enrichment score (NES) that indicates the degree by which a gene set is overrepresented at the extremes of the ranked list (i.e. upregulated or downregulated). Using *fgsea*, we will generate a normalized enrichment score (NES), p-value and Benjamini-Hochberg adjusted p-value for each gene set for each protein. Multiple testing adjustments will be carried out within each gene set category (e.g. canonical pathways, cell type of origin, etc). Gene sets will be considered significant if their adjusted p-value is ≤0.05.

The primary gene sets that will be used to test for enrichment are shown in Table 1 (shown below). We will convert from protein sets to gene sets using the maps provided by SomaLogic (using specifically 'EntrezGeneSymbol' or 'EntrezGeneID'). For protein complexes, we will consider a protein to be contained within a gene set if it includes the product of any gene within that gene set (as per the Discovery analysis plan).

[12] Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst. 2015 Dec 23;1(6):417-425. doi: 10.1016/j.cels.2015.12.004. PMID: 26771021; PMCID: PMC4707969

Table 1: *Gene sets to test for protein enrichment*

Gene set category	***Database***	***Specific gene sets***	***Links***
Pathways / ontologies	*MSigDB*	*KEGG (186 gene sets)*	*http://www.gsea-msigdb.org/gsea/msigdb/genesets.jsp?collection=CP:KEGG*
Pathways / ontologies	*MSigDB*	*GO (10,561 gene sets)*	*http://www.gsea-msigdb.org/gsea/msigdb/genesets.jsp?collection=GO*
Pathways / ontologies	*MSigDB*	*Hallmark (50 gene sets)*	*http://www.gsea-msigdb.org/gsea/msigdb/genesets.jsp?collection=H*
Pathways / ontologies	*MSigDB*	*Reactome (1654 gene sets)*	*https://www.gsea-msigdb.org/gsea/msigdb/human/genesets.jsp?collection=CP:REACTOME*

The primary gene sets we will consider are canonical pathways. We will take these gene sets from the MSig'database (MSigDB), and will include KEGG, Gene Ontology (GO), Hallmark and Reactome gene sets.

Data: Protein lists for each respective clinical outcome per intracellular protein score and non-IPS-adjusted datasets, as generated in subanalysis 2.1 (including estimates, p-values, adjusted p-values etc), will be passed to gene set enrichment analysis. These protein lists will include all proteins examined irrespective of whether a statistically significant association (adjusted p-value of ≤ 0.05) was observed with the given outcome. Clinical outcomes that show no evidence of protein regulation will still be passed through *fgsea,* though these results should be interpreted with caution.

Other Gene Sets / Software for Pathway Analysis Visualisation

Additional gene sets will be explored using plug-ins available in Cytoscape[13] including the '*Human Network*' (HumanConsensusPathDB[14]) gene set accessed through the Phenoscape[15] app menu.

Cytoscape conventionally requires gene abundance data in the form of fold changes and (adjusted) p-values, however, in our case we will instead provide log-odds ratio per standard deviation in protein expression estimates, and adjusted p-values. Cytoscape calculates an abundance score for each node in the network as:

log2foldchange * -log10(pvalue)

Analysed abundance data with gene symbols (i.e. gene names), fold change and P-values are imported into Cytoscape and matched with the loaded network by gene symbol. No thresholding by fold change or P-value is required as "significance is determined at a sub-network level by calculation of empirical P-values through random sampling of the background network"[16].

Results/Plots:

We will visualize these results using bubble plots of significantly differentially expressed gene lists for each gene set category, and we will visualize the protein co-abundance network and enriched pathways using the *RCy3*[17] package in R (alternatively using Cytoscape software directly). Proteins that are statistically significantly regulated (based on adjusted p-values ≤ 0.05) within a differentially enriched pathway will be visualised using *pathview*[18] R package (specifically, for KEGG gene set). We will group our clinical outcomes into two main bubble plots; i) OA-related outcomes, (WOMAC, ordinal KL grade, binary RKOA status, binary advanced RKOS status, PASS) and ii) disease-grouping (comparison of OA vs. DF control proteomes).

---

[13] https://cytoscape.org/.

[14] Kamburov A, Wierling C, Lehrach H, Herwig R. ConsensusPathDB--a database for integrating human functional interaction networks. Nucleic Acids Res. 2009 Jan;37(Database issue):D623-8. doi: 10.1093/nar/gkn698. Epub 2008 Oct 21. PMID: 18940869; PMCID: PMC2686562.

[15] Jamie Soul, Sara L. Dunn, Tim E. Hardingham, Ray P. Boot-Handford, Jean-Marc Schwartz, PhenomeScape: a cytoscape app to identify differentially regulated sub-networks using known disease associations, Bioinformatics, Volume 32, Issue 24, december 2016, Pages 3847–3849, https://doi.org/10.1093/bioinformatics/btw545

[16] Jamie Soul, Sara L. Dunn, Tim E. Hardingham, Ray P. Boot-Handford, Jean-Marc Schwartz, PhenomeScape: a cytoscape app to identify differentially regulated sub-networks using known disease associations, Bioinformatics, Volume 32, Issue 24, december 2016, Pages 3847–3849, https://doi.org/10.1093/bioinformatics/btw545

[17] Gustavsen, A. J, Pai, Shraddha, Isserlin, Ruth, Demchak, Barry, Pico, R. A (2019). "RCy3: Network Biology using Cytoscape from within R." *F1000Research*. doi: 10.12688/f1000research.20887.3.

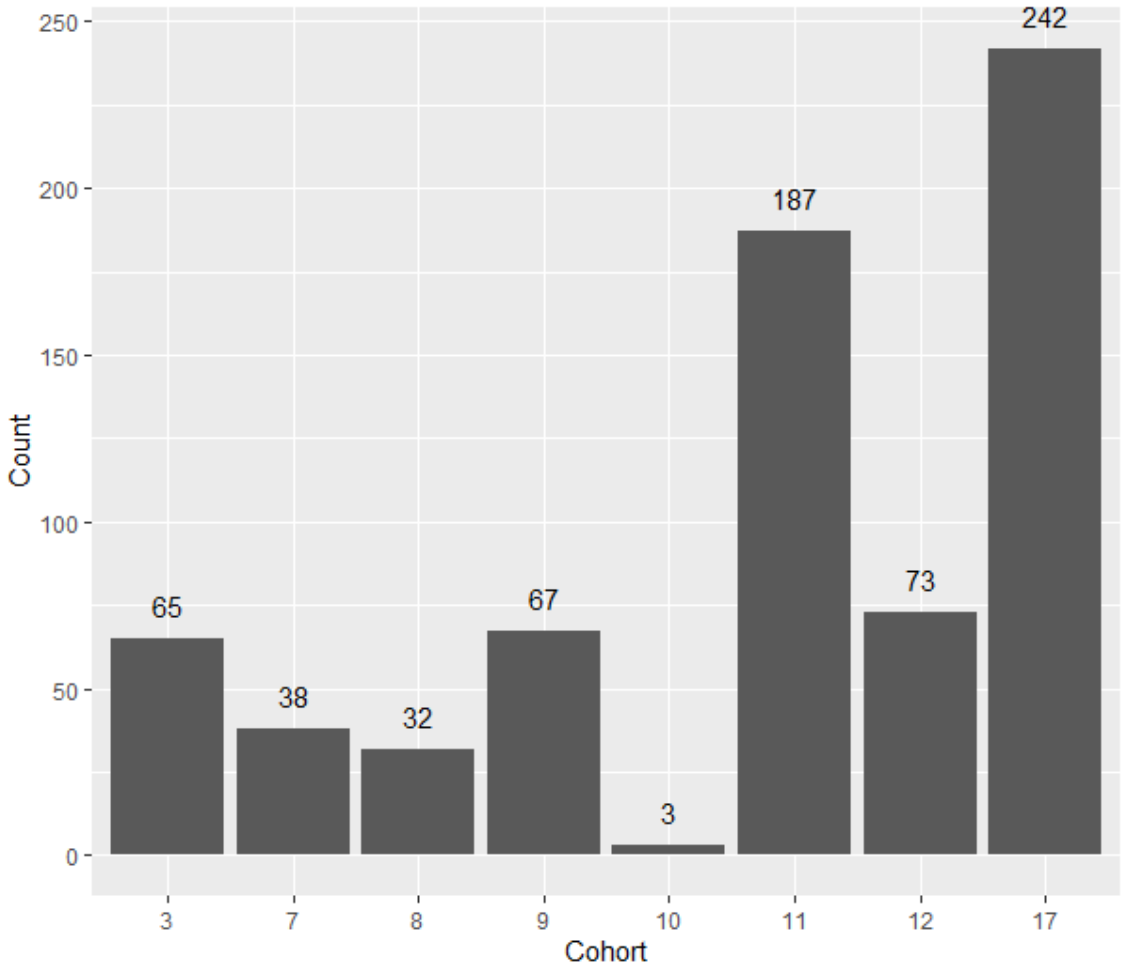
[18] Luo, Weijun, Brouwer, Cory (2013). "Pathview: an R/Bioconductor package for pathway-based data integration and visualization." Bioinformatics, 29(14), 1830-1831. doi: 10.1093/bioinformatics/btt285.

Appendix 1: *Descriptive statistics for the clinical data that comprise the Replication data release*[19]*.*

Note that, due to a database export error, N=6 samples were excluded from data processing and thus from all of the tables and plots in this appendix. Specifically, this was 6 samples of contralateral knees taken at a different time-point from a baseline sample. N=3 of these were samples taken at later visits, and N=3 were samples taken at earlier visits (for the latter, we have assigned the later sample as the baseline sample to ensure we still have data for this individual).

The tables and graphs below give the missingness statistics and the distribution across cohorts for the released phenotype data.

Summary of samples by cohort and tranche

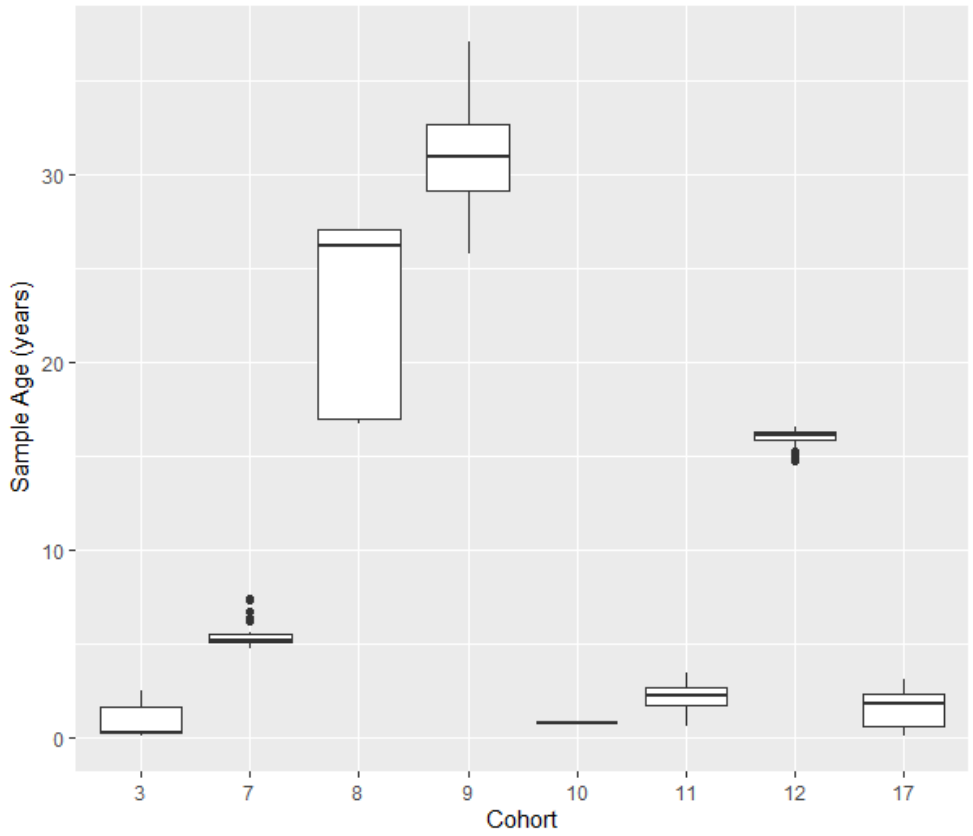


[19] N = 707 samples processed with proteomic data.

Cohort Number	3	7	8	9	10	11	12	17	Total
Sample Count (tranche 3)	55	38	32	67	3	187	73	242	697
Sample Count (tranche 4)	10	-	-	-	-	-	-	-	10
									707[20]

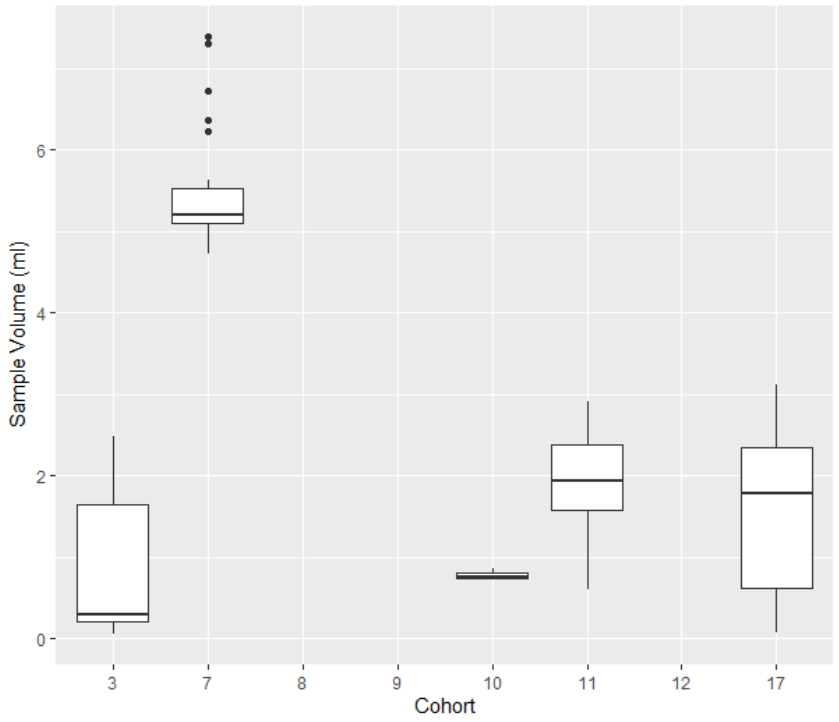
QA Variables[21]:

1. **Age of sample (no missing data)**



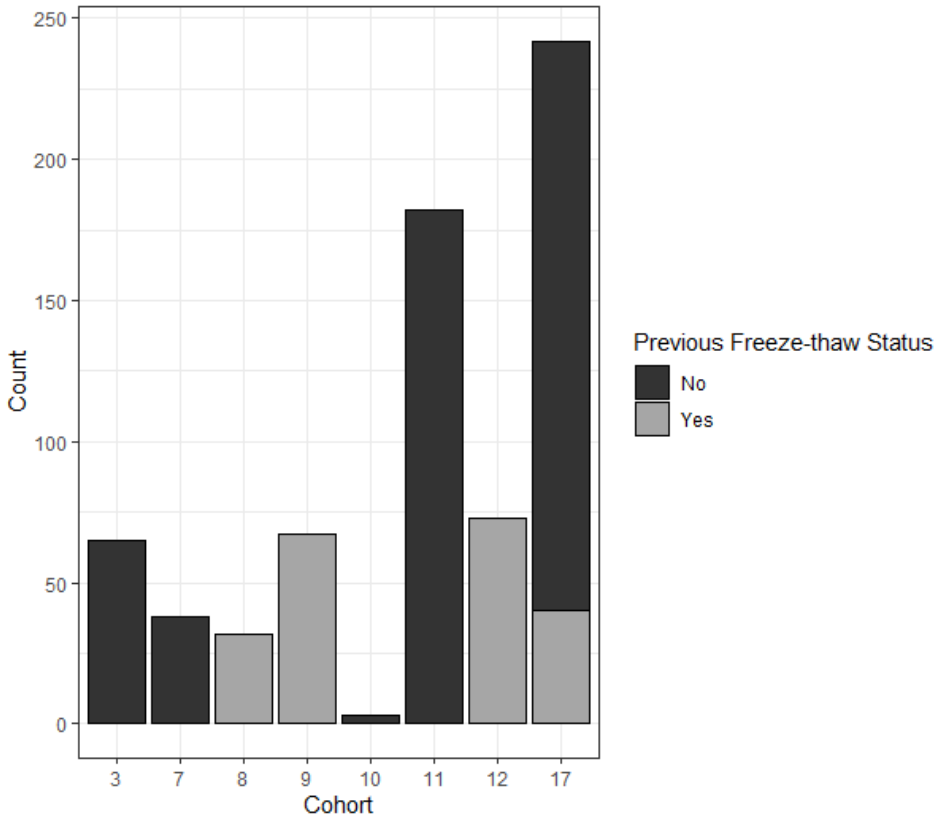
[20]N = 707 samples (from 669 patients) remained. Of these 669 patients, 32 had bilateral sampling at the same study visit so in these cases the right knee was selected for inclusion.A further 6x cases had contralateral sampling at follow-up, for which the right knee was selected for analysis.
[21] Of the N = 707 samples, N = 702 samples had a QC group allocation. All figures and tables relate to N = 702.

2. Sample volume



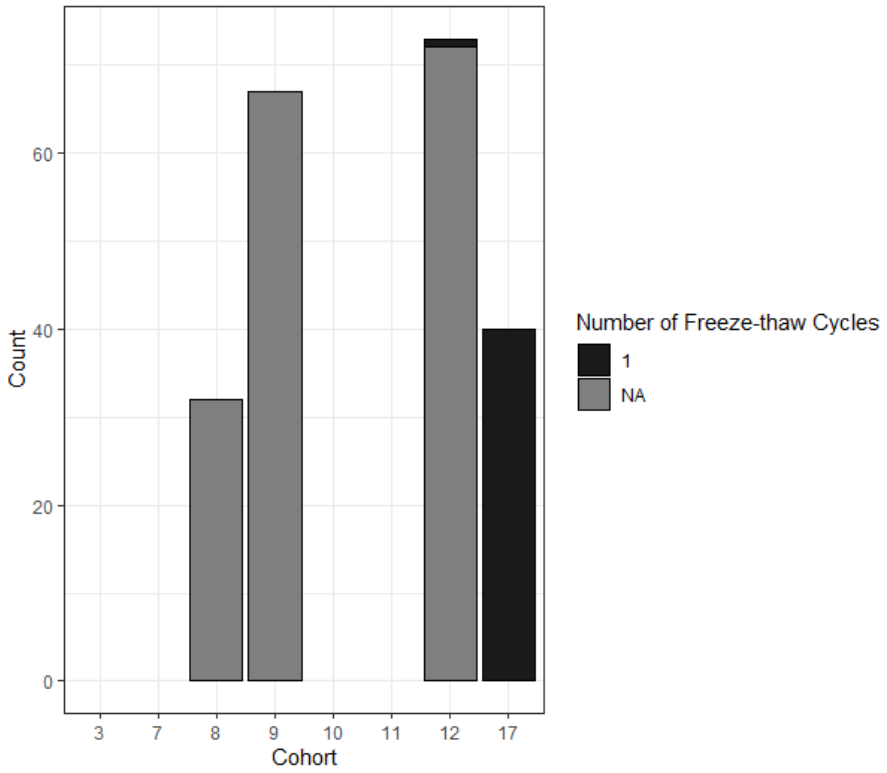
sf_iknee_qc_group	sf_iknee_volume available	missing
702 (OA)	472	230

3. Previous Freeze thaw status



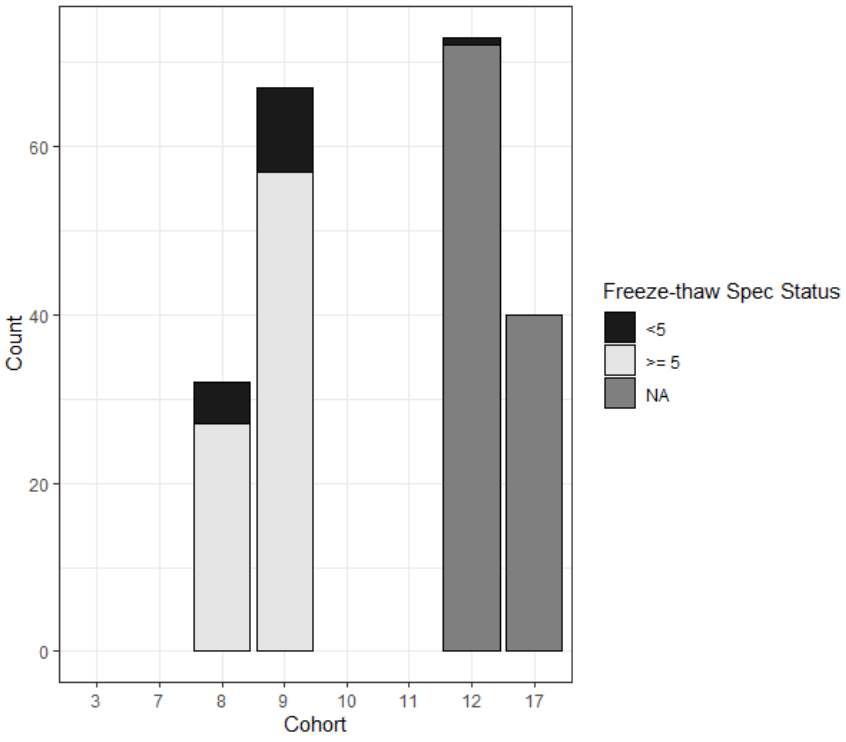
sf_iknee_qc_group	Total number previously freeze-thawed	Total number NOT previously freeze-thawed	missing
702 (OA)	212	490	0

4. **Number of freeze-thaw cycles (for those that underwent at least one previous freeze-thaw)**



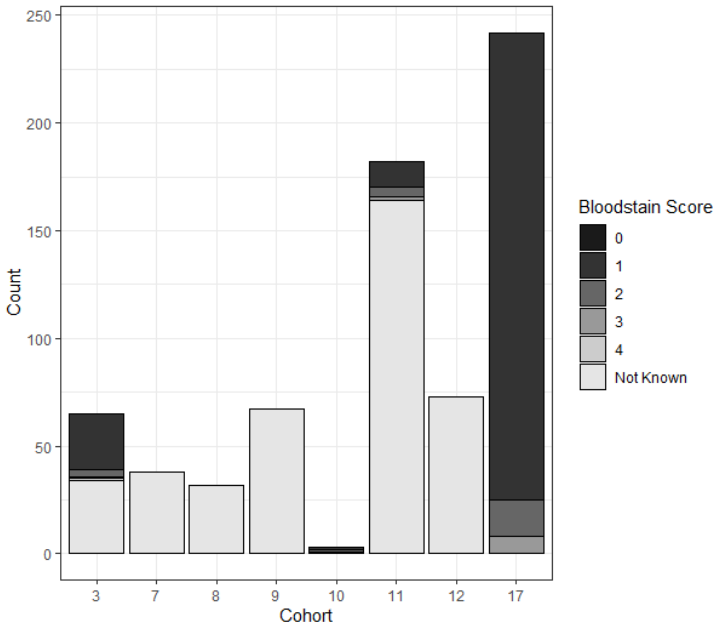
sf_iknee_qc_group	sf_iknee_freezethaw_cycles status Available in samples that previously undergone at least one freeze-thaw	missing
702 (OA)	41	171

5. Five or more freeze-thaws (of samples that underwent previous freeze-thaw)



sf_iknee_qc _group	sf_iknee_freezethaw_spec available in samples previously freeze-thawed	missing
702 (OA)	100	112

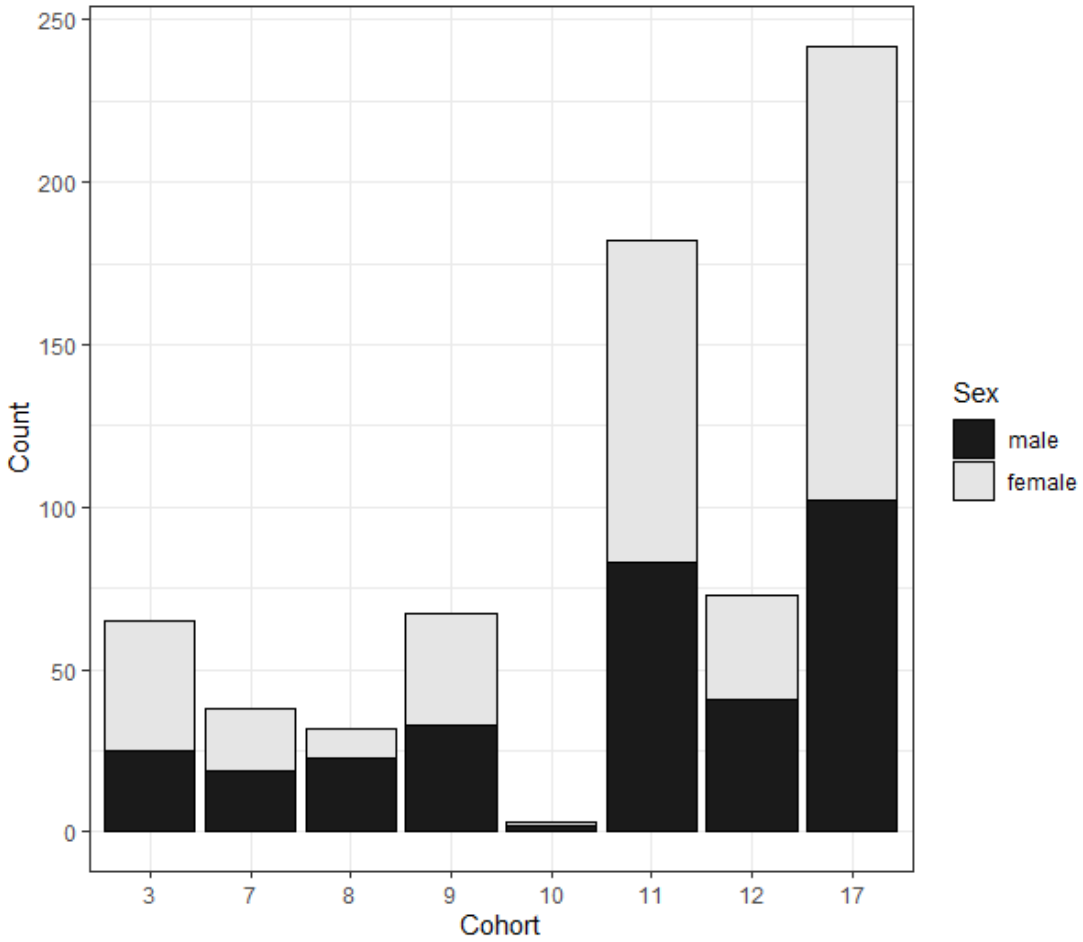
6. Blood Staining



sf_iknee_qc_group	sf_iknee_bloodstaining score available	Missing
702 (OA)	294	408

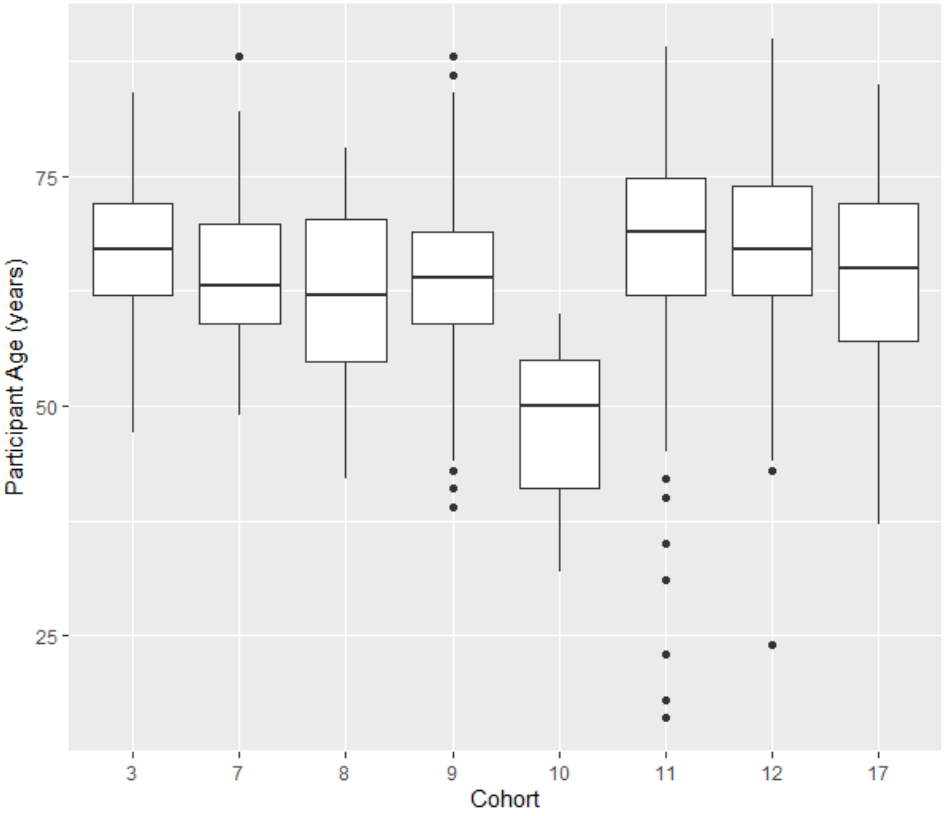
<u>Demographic variables</u>:

1. **<u>Sex</u>**

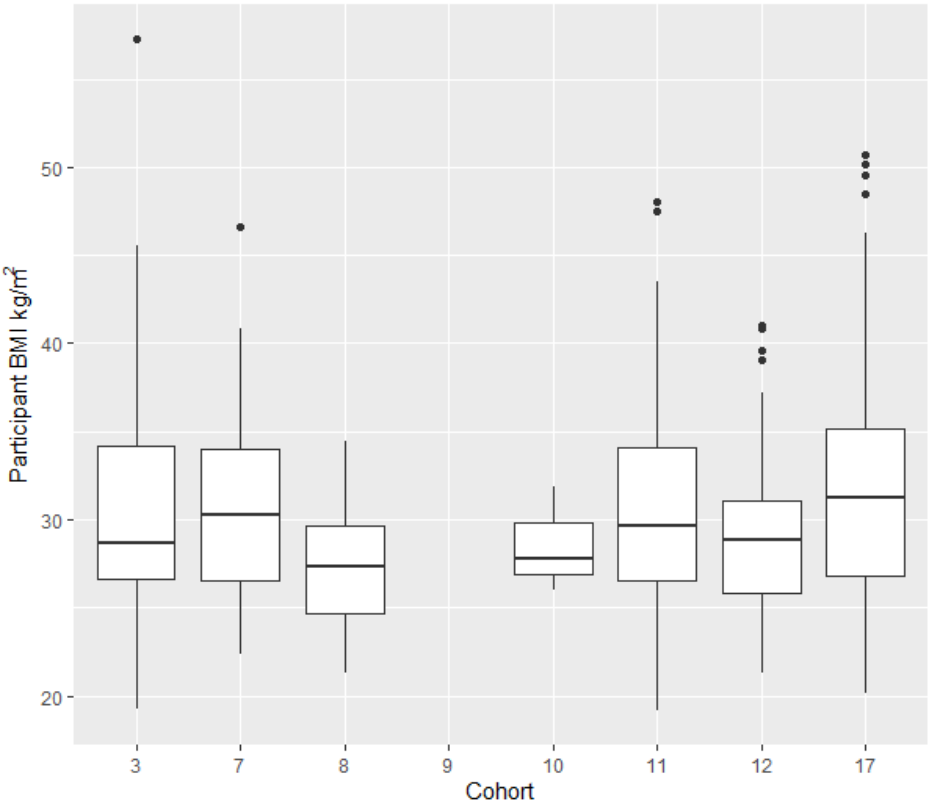


sf_iknee_qc_group	Sex == "m" (male)	Sex == "f" (female)	Missing
702 (OA)	328 (46.7%)	374 (53.3%)	0

2. Participant age (no missing data)

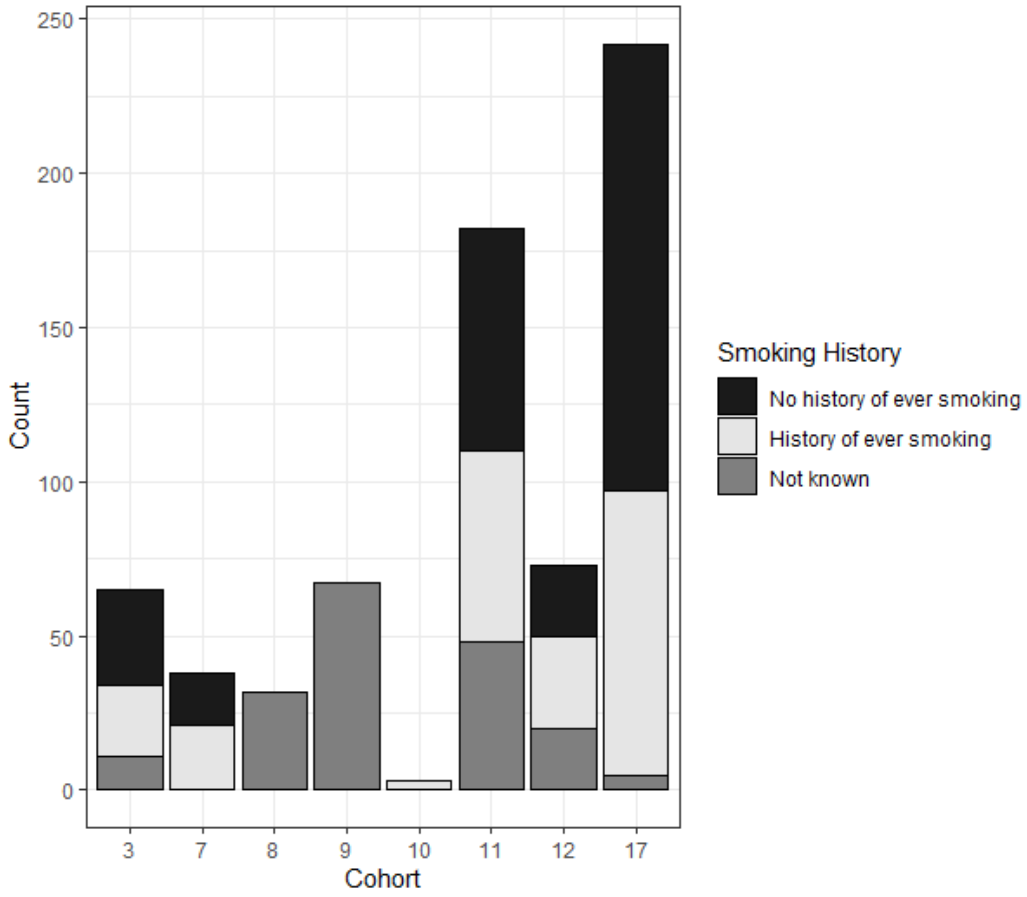


3. Participant BMI



sf_iknee_qc_group	Participant BMI available	Missing
702 (OA)	585	117

4. Smoking History

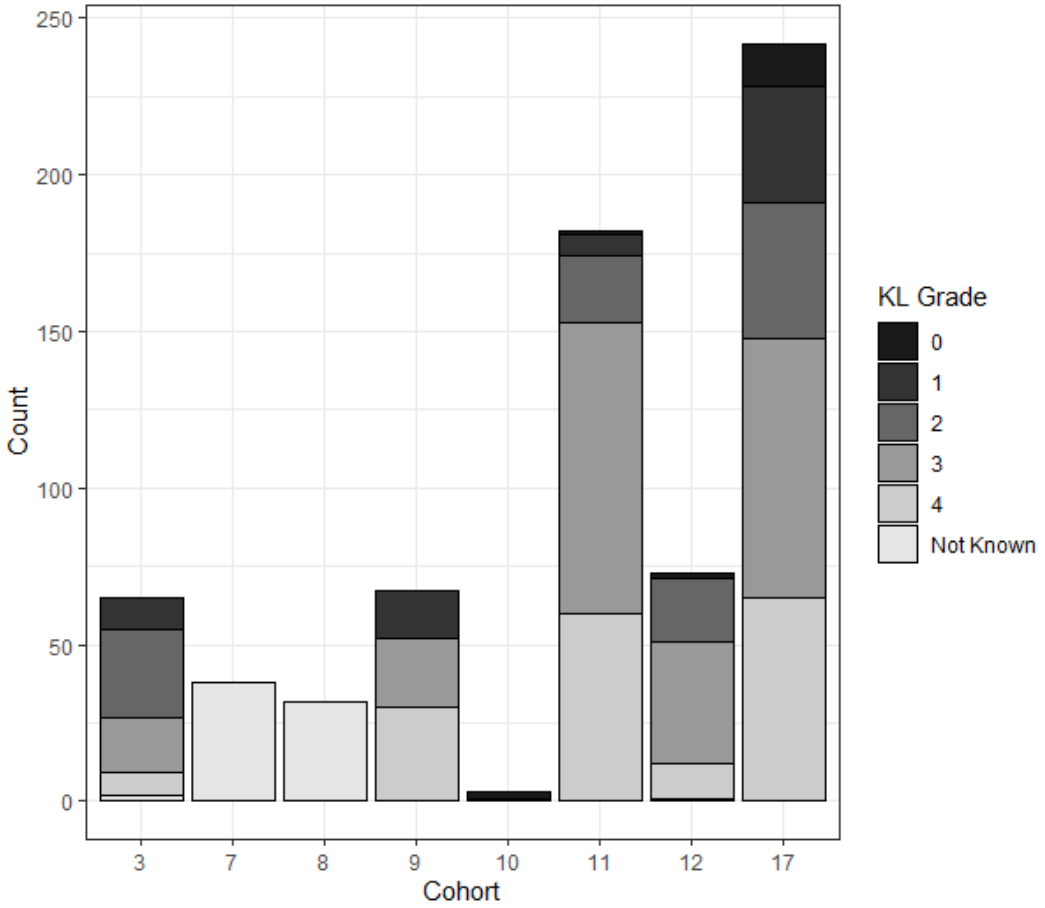


sf_iknee_qc_group	No history of ever smoking	History of ever smoking	Missing
702 (OA)	288 (41.0)	231 (32.9)	183 (26.1)

Radiographic Variables

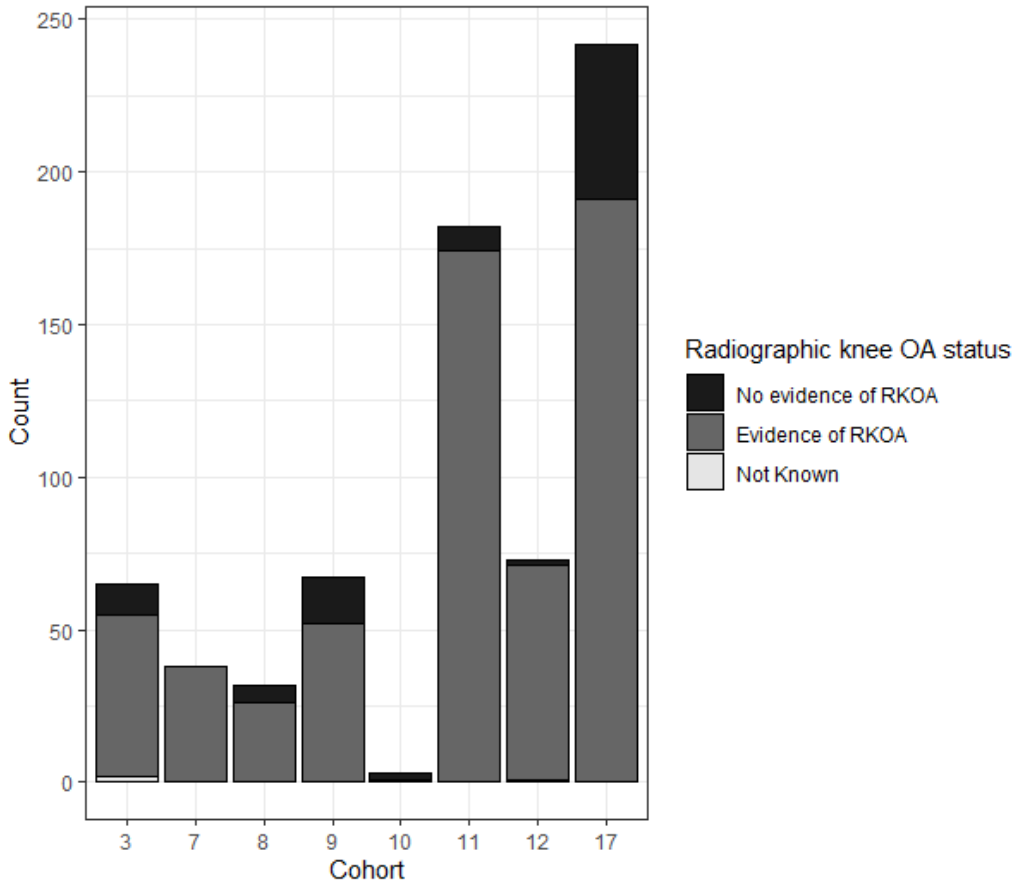
There are 4 samples with no radiographic measure of disease severity (i.e. missing ordinal KL grade, and binary indicators for the presence of radiographic & advanced radiographic knee OA).

1. Ordinal KL Grade



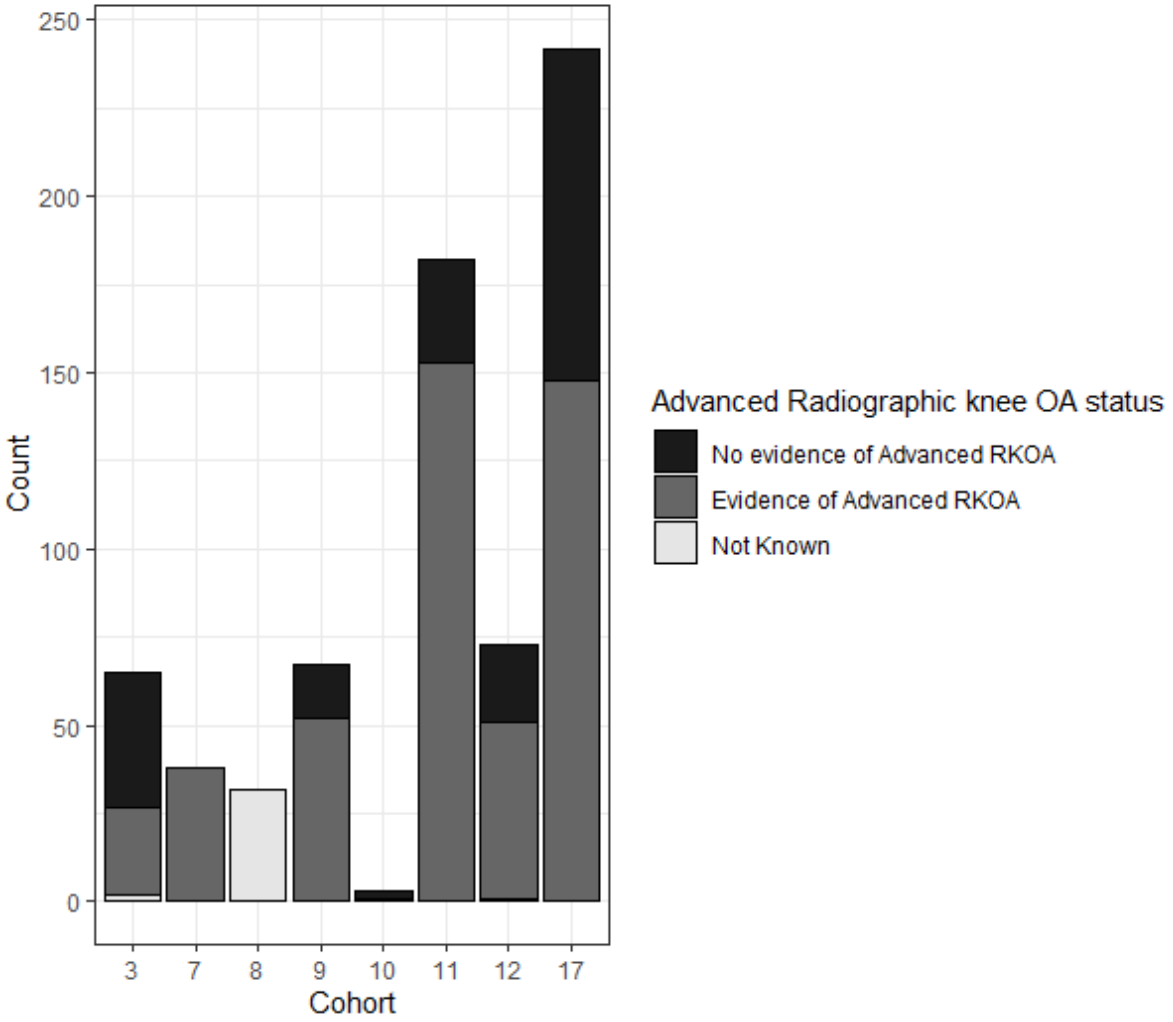
sf_iknee_qc_group	Ordinal KL grade available	Missing
702 (OA)	628	74

2. Radiographic knee OA status



sf_iknee_qc_group	Binary indicator for the presence of radiographic knee OA available	Missing
702 (OA)	698	4

3. Advanced radiographic knee OA status

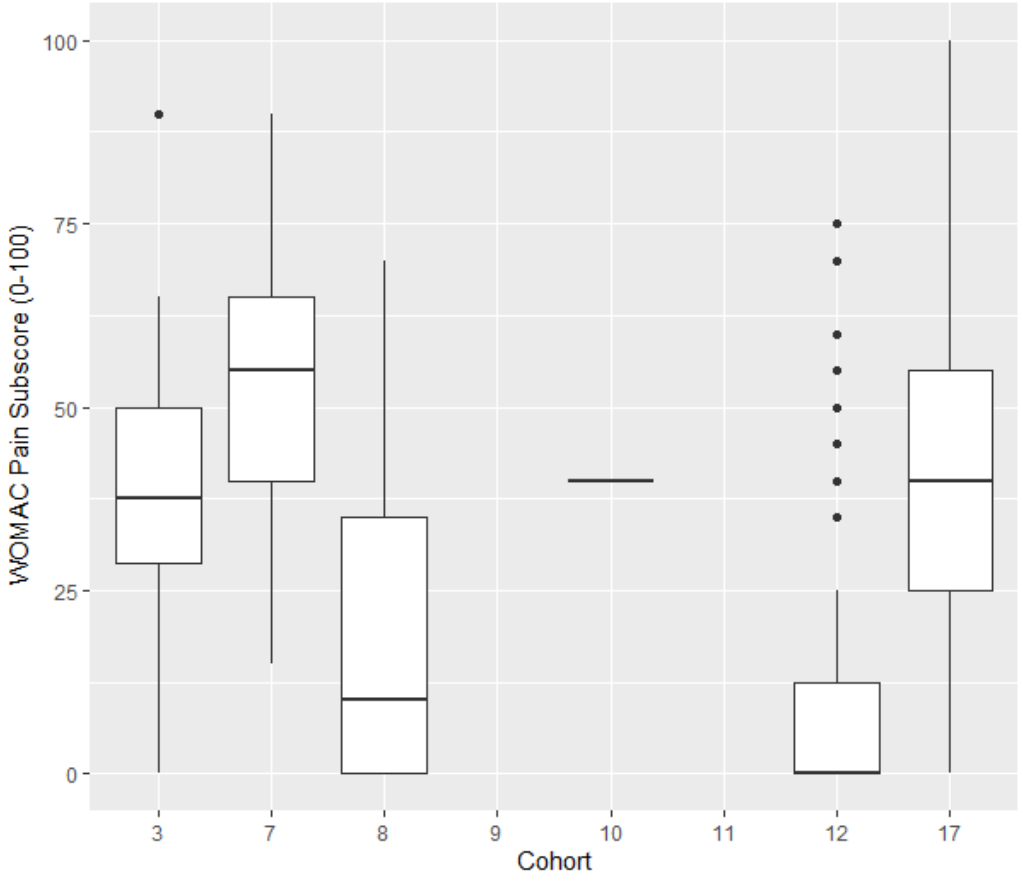


sf_iknee_qc_group	Binary indicator for the presence of advanced radiographic knee OA available	Missing
702 (OA)	666	36

Pain Variables:

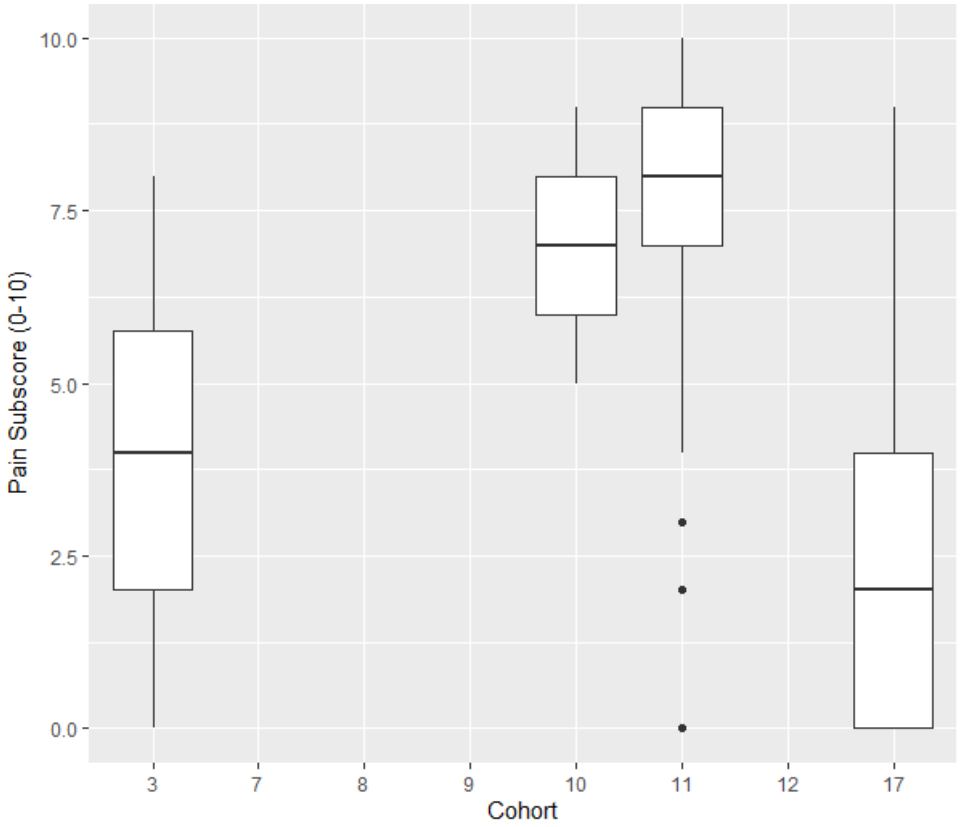
1. **<u>WOMAC pain score (derived from WOMAC or KOOS items) for OA samples</u>**

- Two cohorts do not have WOMAC measures (N = 254)



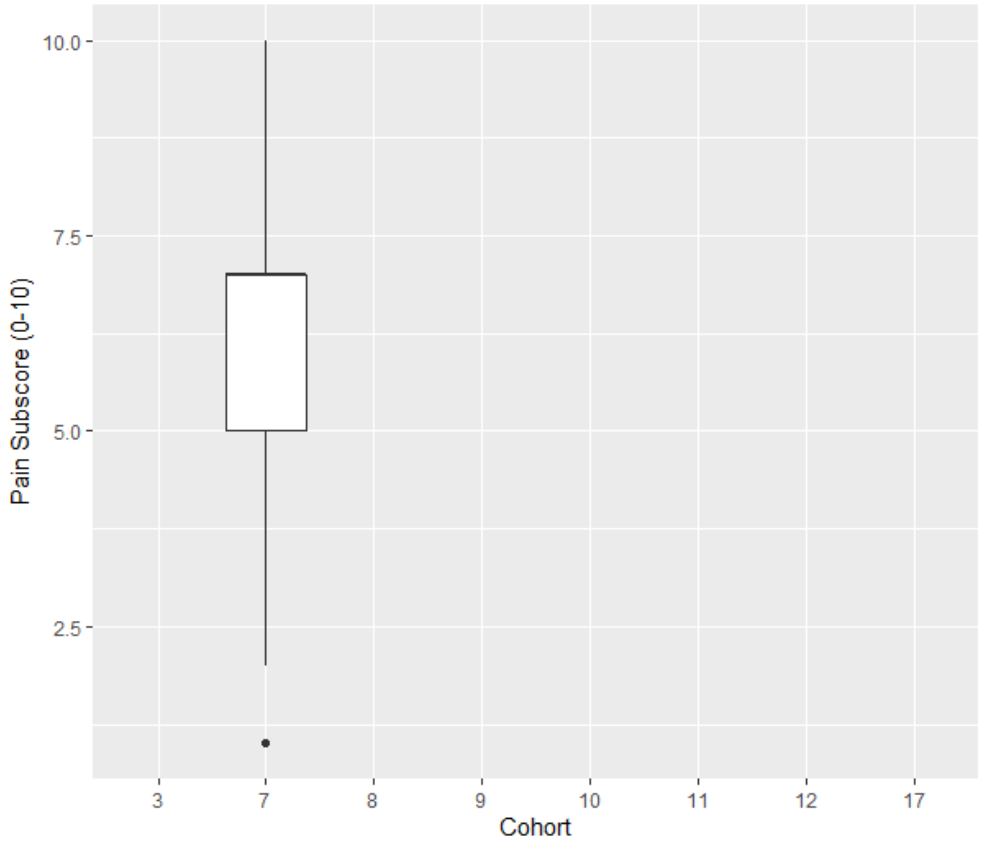
sf_iknee_qc_group	WOMAC Pain Sub-score Available	Missing
702 (OA)	427	275

2. Knee-specific NRS/VAS



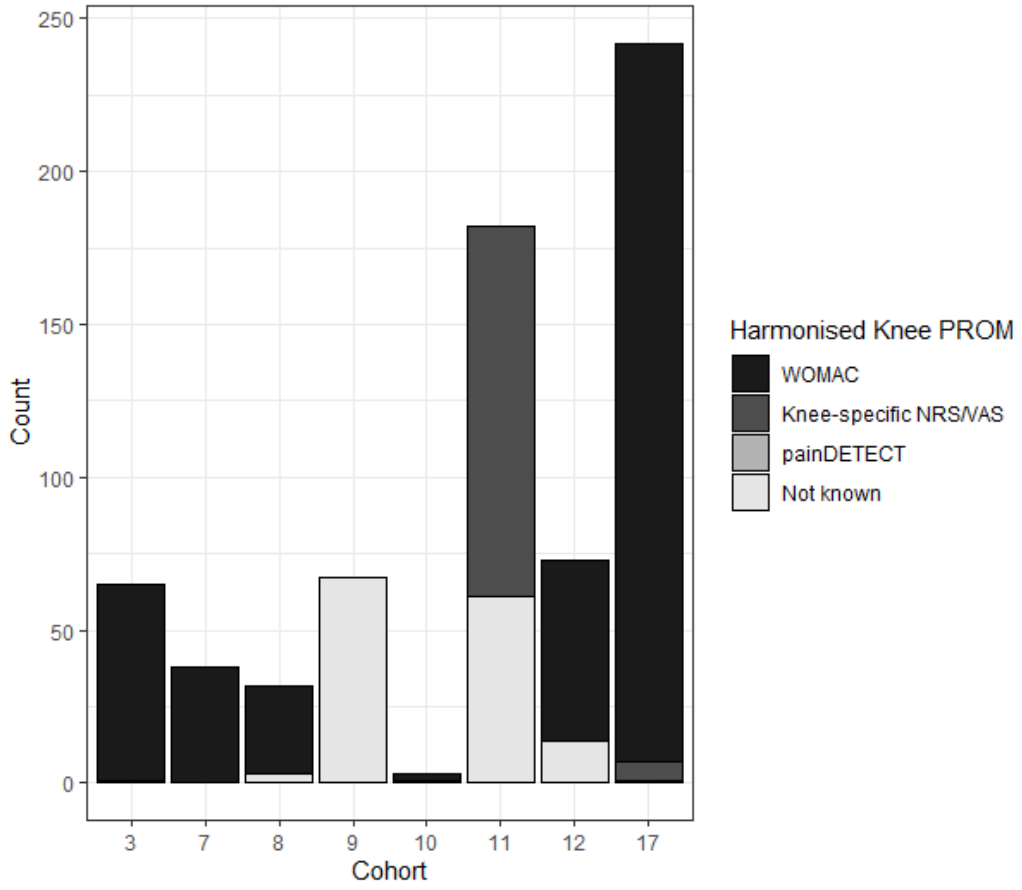
sf_iknee_qc_group	Knee-specific NRS/VAS Available	Missing
702 (OA)	424	278

3. painDETECT (average pain score)



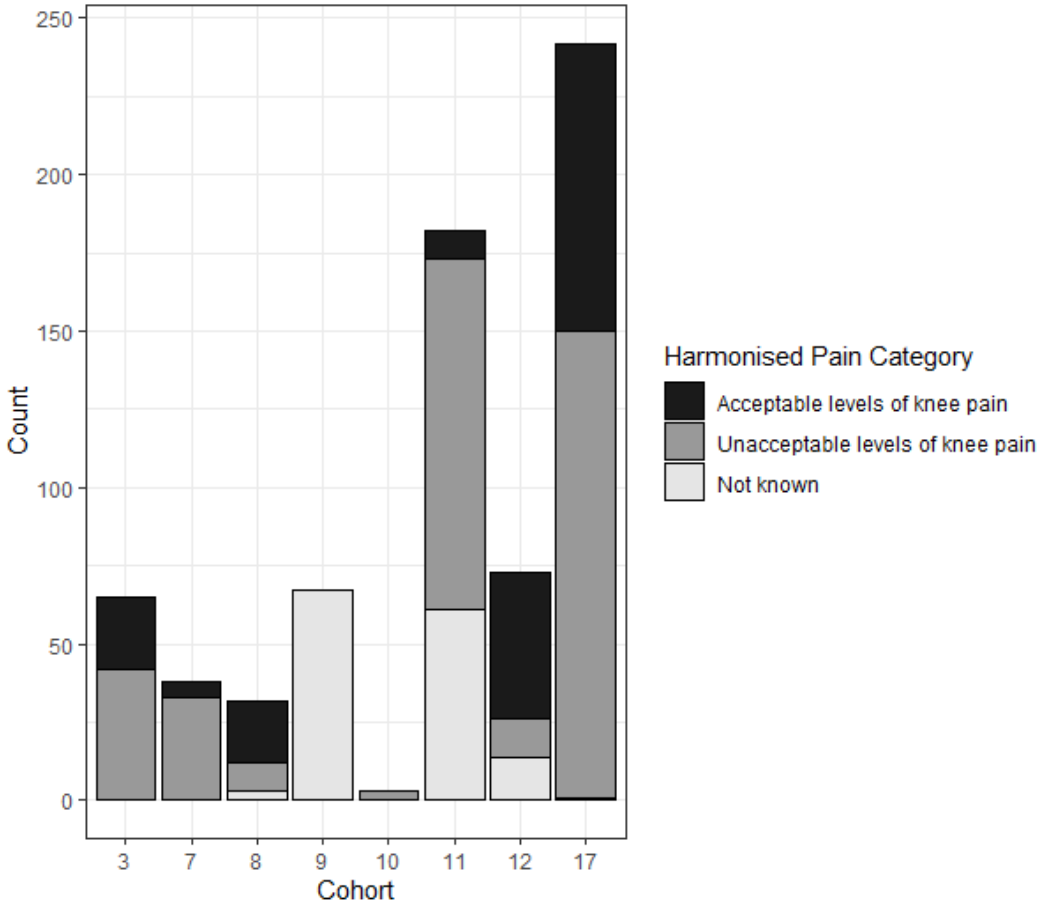
sf_iknee_qc_group	painDETECT Available	Missing
702 (OA)	37	665

4. Harmonised knee patient reported outcome measure (PROM)



sf_iknee_qc_group	Harmonised Knee PROM Available	Missing
702 (OA)	556	146

5. Harmonised Pain Category



sf_iknee_qc_group	Harmonised Pain Category Available	Missing
702 (OA)	556	146

Appendix (2):

replication_QApheno_1: Sample and patient characteristics used in quality control

This dataset includes sample information used to carry out quality assessment on the synovial fluid samples. It includes the following fields:

Field	Description	Coding
sf_iknee_sample_id_number	The STEpUP OA Sample Identification Number (SIN)	string
stepup_id	The STEpUP OA Participant Identification Number (PIN)	string
age_sampling	Patient age at the time sample was taken (to the nearest year)	integer (NA=missing)
sl_plate_id	ID of plate the sample was run on	string
sl_plate_run_date	Date that the same was run	string ("YYYY-MM-DD")
sl_plate_position	Position of the sample on the 96-well plate	string ("XN", where X is row letter and N is the column number)
sl_scanner_id	ID of the scanner that the sample was read using	string
sl_tranche_number	Shipment tranche in which sample was run (tranche3 vs tranche4)	{3 = tranche 3, 4 = tranche 4}
sl_bimodal_signal	The technical bimodal signal, strongly correlated with processing batch, used to batch-correct the data.	{bimodal1, bimodal2 - arbitrary labels for the two groups. NA=missing}
sf_iknee_proc_batch	Batch number for index knee sample	Integer (NA = missing)
sf_iknee_proc_order	Processing order number for index knee sample	Integer (NA = missing)

sf_iknee_proc_treat_date	Date sample was hyaluronidase treated by KIR	Text (dd-mm-yyyy)
sf_iknee_qc_group	Patient grouping (OA, injury or control) at baseline.	{0 = OA, 1 = Joint injury, 2 = healthy control, 3 = inflammatory control, NA = missing}
cohort_name	Cohort ID (an arbitrarily chosen integer assigned to each cohort)	integer
sex	Patient sex at baseline (as defined by individual cohort collectors).	{m = male, f = female, NA = missing}
sample_age	Time between date of sample collection and date of STEpUP OA sample processing for the index knee (years)	float (years) (NA=missing)
sf_iknee_volume	Total SF volume collected (ml)	float (ml)
sf_iknee_prev_freeze_thaw	Has the sample been freeze-thawed prior to STEpUP OA sample processing?	{0 = No, 1 = Yes, NA = Unknown}
sf_iknee_freezethaw_cycles	Number of freeze-thaw cycles (if known)	integer (NA=missing)
sf_iknee_freezethaw_spec	Indicates whether the sample has been freeze-thawed less than, or greater to or equal to five times.	{0 = <5, 1 = ≥5, NA = missing}
sf_iknee_bloodstaining	Grading of SF bloodstaining prior to centrifugation (if known). Scale of 1-4, with larger numbers corresponding to greater degrees of blood staining.	{1 = None , 2 = Mild, 3 = Moderate, 4 = Severe, NA = Not known}
sf_spun_vs_unspun	Indicator for whether the sample was centrifuged prior to receiving at KIR	0 = unspun, 1 = spun, 2 = not known

Appendix (3):

replication_DAPpheno_1: Core clinical phenotype data, excluding pain

This dataset includes the clinical phenotype data required for the analyses above, excluding pain data. It includes the follow fields:

Field	**Description**	**Coding**
sf_iknee_sample_id_number	The STEpUP OA Sample Identification Number (SIN)	string
stepup_id	The STEpUP OA STEpUP Participant Identification Number (PIN)	string
cohort_name	Cohort ID (an arbitrarily chosen integer assigned to each cohort)	integer
sf_iknee_qc_group	Patient grouping (OA, joint injury or control) at baseline. Note that this is a rough description of the patient group based primarily on the inclusion criteria of the individual cohorts, and should not be over-interpreted (e.g. there is no guarantee that the joint injury grouping is OA-free).	{0 = OA, 1 = Joint injury, 2 = healthy control, 3 = inflammatory control, NA = missing}
age_sampling	Patient age at the time sample was taken (to the nearest year)	integer (NA=missing)
sex	Patient sex at baseline (as defined by individual cohort collectors).	{m = male, f = female, NA = missing}
bmi_sampling	Patient body mass index at the time the sample was taken (calculated from provided height and weight or directly provided by cohort collector, in that order of preference)	float (kg/m^2)
kl_grade_worst	Ordinal Kellgren-Lawrence grade of radiographic severity at time of sampling.	{0 = grade 0 (none), 1 = grade 1 (doubtful), 2 = grade 2 (minimal), 3 = grade 3 (moderate), 4 = grade 4 (severe), NA = Missing OR Not Known}
radiographic_knee_oa	Flag indicating whether the sample was taken from a patient with radiographic OA in the index knee, defined as a KL grade greater or equal to two at time of sampling.	{0 = No (i.e. KL < 2), 1 = Yes (i.e. KL >= 2), NA = Missing

		OR Not Known}
kl_grade_advanced	Flag indicating whether the sample was taken from a patient with advanced radiographic OA in the index knee, defined as a KL grade greater or equal to three at time of sampling.	{0 = No (i.e. KL < 3), 1 = Yes (i.e. KL >= 3), NA = Missing OR Not Known}
smoking_history	Flag indicating whether the patient was a current or past smoker at the time of the baseline sample.	{0 = No (i.e. never smoked), 1 = Yes (i.e. current smoker or past smoker), NA = missing or not available}
baseline	Flag indicating whether this sample is a baseline sample (as defined in the Introduction) or is the primary sample to be included in analysis (right knee to be used in cases of bilateral sampling at the same visit).	{0 = No, 1 = Yes}

replication_DAPpheno_2: Core pain phenotype data

This dataset includes the continuous and binary patient-reported pain data required for the analyses above. The release includes the follow fields:

Field	**Description**	**Coding**
sf_iknee_sample_id_number	The STEpUP OA Sample Identification Number (SIN)	string
stepup_id	The STEpUP OA Participant Identification Number (PIN)	string
harm_knee_pain	Binary flag indicating whether experienced pain is above the Patient Acceptable Symptom State (PASS) at the time of sampling (calculated manually from the KOOS pain subscale, the WOMAC pain subscale or knee VAS (knee-specific NRS/VAS or painDETECT VAS, in order of preference). Yes vs No.	{0 = No (acceptable pain), 1 = Yes (unacceptable pain), NA = missing or Not Available.}
harm_pain_prom	The specific patient reported outcome measure used to derive harm_knee_pain.	{1 = KOOS, 2 = WOMAC, 3 = Knee specific

		VAS/NRS, 4 = PainDETECT VAS, NA=missing}
koos_pain	KOOS pain subscore (calculated from full KOOS questionnaire results, or from combined subscore provided by cohort collectors, in that order of preference). Scale of 0-100, where 0 is the worst possible pain recordable.	float
womac_pain	WOMAC pain subscore (calculated from full WOMAC questionnaire results, or from combined subscore provided by cohort collectors, or derived from full KOOS questionnaire results, in that order of preference). Scale of 0-100, where 100 is the worst possible pain recordable.	integer
knee_pain_nrs	Patient reported knee pain on a Numeric Rating Scale (0-10), where 10 is the worst pain imaginable.	float
pd_pain_average_score	Patient reported average pain score (over the last 4 weeks) from the painDETECT questionnaire. Scale of 0-10, where 10 is the worst pain imaginable.	integer

Appendix (4):

Four datasets will be included in the replication data release

- Plate/Bimodal Batch Corrected
- IPS adjusted + Plate/Bimodal Corrected
- Plate/Bimodal/Spinning Batch Corrected
- IPS adjusted + Plate/Bimodal/Spinning Batch Corrected

The release also includes filter files to indicate samples and proteins removed using the filters described in the QA report. Note that protein concentration data is included in the protein concentration files for all proteins regardless of filter status. A summary of the number of samples and proteins removed by each filter is included below:

Filter label in file	**Filter name**	**Description**	**Applies to**	**IPS adjusted + Plate/ Bimodal/ Spinning Batch Corrected**	**Plate/ Bimodal/ Spinning Batch Corrected**	**IPS adjusted + Plate/ Bimodal Batch Corrected**	**Plate/ Bimodal Batch Corrected**

NONHUMAN	Non-human proteins	Non-human or control proteins	Proteins	307	307	307	307
OA_REPO	Reproducibility in OA pool	Predicted R2 < 0.5	Proteins	485	485	485	485
INJ_REPO	Reproducibility in injury pool	Predicted R2 < 0.5	Proteins	252	252	252	252
FREEZETHAW_CONFOUND	Associated with number of freeze-thaw cycles	ANOVA p < 0.05/7289 (conditional on cohort)	Proteins	254	60	212	56
SAMPLEAGE_CONFOUND	Associated with sample age	ANOVA p < 0.05/7289 (conditional on cohort)	Proteins	169	97	229	77
BIMODAL_CONFOUND	Associated with bimodal signal	ANOVA p < 0.05/7289	Proteins	359	348	72	96
SPINNING_CONFOUND	Associated with sample spinning status	p < 0.05 Pearson correlation test on 18 paired spun/unspun samples	proteins	844	844	0	0
SOMASCAN_FAIL	SomaLogic inhouse QC	Hybridization Scale Factor > 2.5	Samples	1	1	1	1
LOD_SAMPLE	Limit of detection	25% of proteins below/above limit of detection	Samples	3	3	3	3
TOTPROT_OUTLIER	Total protein outliers	>5SDs from mean	Samples	4	4	4	4
PCA_OUTLIER	PCA outliers	>5SD from combined centre on top PCs	Samples	7	7	7	7
	Total remaining		**Proteins Samples**	**5404/7596 690/701**	**5626/7596 690/701**	**6290/7596 690/701**	**6558/7596 690/701**