STEpUP OA discovery analysis plan

v1.0, 29/07/2022

Introduction 2

Quality control process 2

Data analysis plan 3

Overview of analysis approach 3

Primary analysis: Synovial fluid endotypes in OA 4

Sub-analysis 1.1: Unsupervised clustering 4

Sub-analysis 1.2: Identifying characteristic proteins of clusters 6

Sub-analysis 1.3: Bioinformatic characterisation of clusters 7

Sub-analysis 1.4: Clinical characteristics of endotypes 9

Sub-analysis 1.5: Correlation of technical confounders with endotypes 11

Secondary analysis: Synovial proteomics of OA clinical features 11

Sub-analysis 2.1: Finding SF correlates of clinical features 11

Sub-analysis 2.2: Bioinformatic characterisation of clinical features 12

Network analysis: Co-expression and regulation of synovial fluid proteins in OA and injury 13

Sub-analysis 3.1: Co-expression analysis 13

Sub-analysis 3.2: Investigating drivers of synovial protein co-expression 14

Appendix 1: Data releases associated with this data analysis plan 16

discovery_somascan_1: Proteomic quantification, normalization, batch correction and QC 16

discovery_QApheno_1: Sample and patient characteristics used in quality control 17

discovery_DAPpheno_1: Core clinical phenotype data, excluding pain 18

discovery_DAPpheno_2: Core pain phenotype data 19

Appendix 2: Descriptive statistics for the clinical data 21

QA variables 21

Age of sample (no missing data) 21

Demographic variables 25

Radiographic variables 28

Pain variables 30

Appendix 3: Power calculations 35

Appendix 4: Lay summary of the analysis plan 38

Introduction

This document outlines the plan for the analysis of the STEpUP OA discovery cohort. The dataset for this analysis consists of:

1. SomaScan data on a total of 1045 synovial fluid samples, split across two tranches and 22 plates. This includes baseline samples from the knee joints of osteoarthritis (OA, N=719) and joint injury (N=218) patients, with N=64 follow-up samples from later visits (for 61 total patients), as well as with small numbers of inflammatory arthritis (N=5) and healthy control (N=37) samples, along with two samples with missing diagnosis information. A breakdown of these samples by cohort, by sample type and by tranche are summarized in the following tables.

Sample Type Breakdown

Tranche	Baseline OA	Baseline Injury	Repeated samples		Healthy control	Inflammatory arthritis control
			OA	Injury		
Tranche1	257	174	0	0	2	0
Tranche2	462	44	36	28	35	5
Tranche 1&2	719	218	36	28	37	5

2. A basic clinical dataset. All samples have basic demographics (age, sex, cohort number). The majority of patients also have pain data (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 (e.g. Kellgren-Lawrence scores) available. We also have data available on confounders (BMI and smoking history) for a subset of samples.

Details of the datasets and fields that are available for this data analysis are given in Appendix 1. Descriptive statistics for each disease group and cohort are given in Appendix 2.

The primary analysis set for this analysis plan will consist of the baseline OA and injury samples. 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.

Quality control process

Normalization and quality assessment of the SomaScan data will be carried out as described in the STEpUP OA QA Report v2.0. In summary, there are two proteomic data releases associated with this analysis plan:

A. The primary dataset, which consists of batch-corrected relative concentration data (corrected using ComBat to remove the effects of plate and a further identified technical-variable-associated bimodal signal identified on the PCA analysis)
B. A secondary dataset, consisting of non-batch-corrected relative concentration data, which is used for robustness analyses.

These two datasets represent two different approaches to dealing with technical variation. The primary dataset, A, is less heavily filtered and thus contains more proteins (5944), and has technical variation removed via batch-correction on each protein. The secondary dataset is more heavily filtered, including fewer proteins (4734), and may include residual effects of technical variation, but has not undergone a more dramatic transformation due to batch-correction. Note that each dataset has first undergone standard normalization (including removal of background signal, plate scaling and calibration using plasma calibrators), as established during the past tranche 1 QA analysis.

Samples and proteins are filtered out based on performance metrics (as described in the QA report), as well as filtering down to only human proteins, and removing all control proteins and SOMAmers.

The QA process was used to generate an imputed blood grade (a measure of blood contamination in each sample) for all samples, based on blood-specific protein biomarkers. After further testing we found that multiple markers did not add significant accuracy to the imputation, so “imputed” blood staining grade will simply be the log concentration of a single marker of haemoglobin (HBA1.HBB.4915.64). In all cases below where we correct for blood staining, we will also correct for haemoglobin concentration as a separate robustness analysis.

Data analysis plan

Overview of analysis approach

The analysis plan below is broken down into three sections: a primary analysis, a secondary analysis, and a tertiary network analysis. The analyses are organised around the major questions that they intend to answer.These analyses overlap, and in some cases outputs of the secondary or tertiary analyses will be used to address questions in the primary analysis.

The primary analysis is designed to answer the key consortium question “Are there multiple synovial fluid molecular endotypes in OA?”. The sub-analyses within this section aim to answer the question using unsupervised clustering and provide a bioinformatic characterisation of the endotypes/clusters discovered.

The secondary analysis is designed to build a reference set of proteins and pathways that are associated with clinical parameters of OA and joint injury in synovial fluid. This is independent of the primary aim and will ensure that the STEpUP OA consortium members and the wider community can make maximum use of the data.

The network analysis is designed to study the co-expression of proteins in synovial fluid in different disease groups. The aim of this analysis is to investigate the extent of protein co-expression in synovial fluid, to describe the functional characteristics of these groups of co-expression proteins, and to assess how these differ across different groups of patients. This analysis provides important context for the findings of the primary and secondary analyses, by placing discovered proteins in the wider context of protein co-expression and other protein networks.

Each analysis is broken down into sub-analyses. In the analysis plan below we give the questions that each sub-analysis is designed to answer, the output that we will generate to answer these questions, and a description of the analysis approach that we will take. The secondary sub-analyses 2.1 and 2.2 use a very similar methodology to primary sub-analyses 1.2 and 1.3, so mostly refer back to these sections.

For all analyses we will (unless otherwise specified):

- only include the baseline sample
- carry out separate analyses for joint injury and OA samples
- remove samples that have missing data in the specific clinical or demographic data variables under analysis, where relevant
- be carried out on the primary normalised dataset described above, then repeated on the secondary dataset as a robustness test, with similarities and differences between the two analyses reported on.

To aid interpretation of the results of these analyses, power calculations for major sub-analyses (including the detection of endotypes and the detection of correlations between protein levels and clinical characteristics) are given in Appendix 3.

Throughout the document, all analyses are assumed to be carried out in R[1] unless otherwise specified.

A lay description of these analyses, for the purposes of patient and public involvement, is given in Appendix 4.

Note that there will also be a separate replication analysis plan, designed based on the results of the discovery analysis. There will also be a separate joint injury analysis plan. Analysis of longitudinal data will also be included in a subsequent analysis plan(s).

Primary analysis: Synovial fluid endotypes in OA

Primary analysis questions: Are there multiple synovial fluid molecular endotypes in OA? What are the biological functions of proteins involved in each endotype? Do these endotypes correlate with clinical features?

Sub-analysis 1.1: Unsupervised clustering

Questions:

- How are the proteomic data clustered in OA and in injury? How many clusters are there in each group?

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

    - Does this clustering remain if all samples (OA and injury) are clustered together?
- Are these data well described by multiple clusters?
    - Is the number of clusters statistically significantly different from 1?
    - Is this data well described by clustered as opposed to continuous non-clustered variation?

Results:

- Principal components for each samples (separately within OA and joint injury, and combined)
- UMAP coordinates for each sample (separately within OA and joint injury, and combined)
- Cluster assignments for each sample (separately within OA and joint injury, and combined)
- The value of each of the clustering criteria metrics for each possible number of clusters
- The value of each of the clustering criteria metrics for optimal number of clusters

Plots:

- Plots of the clustering metrics against cluster numbers.
- Plot of clusters by principal components and UMAP coordinates.

Method:

We will carry out dimensional reduction on batch-corrected QC+ normalised log protein quantification using unscaled PCA, and will select the top principal components explaining 80% variation to construct the reduced feature space. We will carry out unsupervised clustering on this reduced feature space using k-means clustering with 10 sets of random starting values.

We will assess clustering primarily using the f(K) statistic, which we will visualise across cluster numbers, and will judge the data to be significantly clustered for any K with f(K)>0.8 5[2] . If there is no cluster number K > 1 resulting in f(K)>0.85, we will conclude that we do not have enough evidence to detect a strong clustering structure. We will also produce plots of the other three popular metrics (silhouette score, gap statistic and elbow methods) against cluster numbers to test the robustness of the conclusion.

If the data are clustered, we will pick the cluster number by majority vote across different clustering metrics (as implemented in the R package *NbClust*[3]) for downstream analyses. We will check the clustering for the selected value of K by visual inspection of the PCA and UMAP plots.

We will also use alternative parameterisation (different clustering algorithms, including the algorithm implemented in the R package *cclust*[4], different parameter initialisations), as well as sampling-with-replacement, to assess the robustness of the clustering and downstream conclusions. We will also test different approaches to feature engineering and dimensional reduction, including carrying out kmeans clustering on eigenprotein values generated from the WGCNA analysis in Sub-analysis 3.1, and using sparse clustering directly on all measured proteins using the R package *sparscl*[5]. In each case, robustness will be measured by

[2] https://www.ee.columbia.edu/~dpwe/papers/PhamDN05-kmeans.pdf

[3] https://cran.r-project.org/web/packages/NbClust/index.html

[4] https://cran.r-project.org/web/packages/cclust/index.html

[5] https://cran.r-project.org/web/packages/sparcl/index.html

estimating the adjusted Rand index between the various alternative clustering techniques and the original (PCA, kmeans-based) clustering.

As with all analysis, the clustering mentioned above will be carried out on the primary batch-corrected protein concentrations data, with a robustness analysis on the secondary non-batch-corrected protein concentration data.

Our primary clustering analysis (and all follow-on analyses below) will be carried out separately within the joint injury samples and OA samples separately, but we will also carry out an all-sample clustering to test how well these clusters generalise between injury and OA. All comparisons will use the adjusted Rand index.

Sub-analysis 1.2: Identifying characteristic proteins of clusters

Questions:

- Which specific protein signatures distinguish between the clusters?
- Do these proteins fall into correlated sets, or are they independent?

Results:

- A table of differential abundance test statistics (p-value and odds ratio) for each protein for each cluster, with separate results conditioned and not conditioned on blood staining
- A table of co-expressed protein modules that associate with endotype.
- Grouping of signature proteins into co-expressed modules.

Plots:

- Violin plot of expression levels of signature proteins for different clusters.
- Heatmap showing associations between protein expression and endotypes.

Methods:

We will test which proteins are differentially expressed in each of the clusters using logistic regression, where the independent variable is protein concentration and the binary outcome variable for each sample is 1 if that sample belongs to the tested cluster and 0 if not. Benjamini-Hochberg will be used to adjust for multiple testing within each cluster. Proteins with adjusted p-values < 0.05 will be reported as signature proteins for the enquiry endotype. Note that these p-values will not be uniformly distributed under the null, and thus should not be considered as true p-values but only as a measure of the relative strength of the differential expression for each protein. We will carry out a robustness analysis to see whether these signals are driven by the presence of blood in the sample, by including imputed blood staining grade as a covariate in the logistic regression and test whether this produces similar association results to the non-conditioned analysis. We will also test robustness of these lists by calculating the overlap with proteins with non-zero coefficients in the sparse clustering results from Sub-analysis 1.1

If more than two clusters are identified, we will also carry out a multinomial regression analysis, with p-values calculated using a likelihood ratio test, in order to increase power to identify proteins that are differentially expressed in a particular subset of clusters.

We will use the weighted gene correlation network analysis (WGCNA[6]) described in sub-analysis 3.1 below to group the signature genes discovered above into co-expressed modules. We will also test each discovered co-expressed module (represented by its eigengene) for correlation with endotype using logistic regression (testing against a p-value threshold of 0.05/Nclusters).

Sub-analysis 1.3: Bioinformatic characterisation of clusters

Questions:

- What are the functional/biological features that distinguish between the clusters?
  - Which canonical pathways or functions are enriched in the cluster-distinguishing protein lists?
  - Are individual clusters distinguished by proteins characteristic of a particular set of cell types?
  - Do the proteins that distinguish each cluster localize to a particular part of the cell?
  - Which upstream modulators can explain differences in protein levels between the clusters?
  - Are any clusters enriched for heritability in genome-wide association studies (GWAS) of OA?

Results:

- 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.
- Partitioned heritability estimates for each cluster based on OA GWAS.

Plots:

- Bubble charts of enriched gene sets (pathways, cell/tissue types, etc) for each endotype
- Bar plots of cell type composition and subcellular type composition of the signature proteins of each cluster.
- Hierarchically clustered heatmaps of the most highly differentially expressed genes, with colour bars showing which key pathways each protein belongs to
- Network plot of proteins involved in enriched pathways.

Methods:

The overall approach in this section will be to test for enrichment of the per-cluster differential expression (calculated in sub-analysis 1.2) in a variety of gene sets representing canonical pathways, cell type of origin, and subcellular localization. Protein set enrichment testing will be performed using the *fgsea* package, ranking proteins by p-value. This will generate a normalized enrichment score (NES), p-value and Benjamini-Hochberg adjusted p-value for each gene set for each cluster. Rank-based testing does not explicitly assume a background protein set, but it will implicitly assume the background (i.e. the universe of proteins) is all proteins tested by SomaLogic that passed QC.

[6] https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-9-559

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 less than 0.05.

The gene sets that we will test for are shown in Table 1 below. We will convert from protein sets to gene sets using the maps provided by SomaLogic. 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.

Gene set category	**Database**	**Specific gene sets**	**Link**
Pathways/ontologies	MSigDB	KEGG	http://www.gsea-msigdb.org/gsea/msigdb/genesets.jsp?collection=CP:KEGG
Pathways/ontologies	MSigDB	GO	http://www.gsea-msigdb.org/gsea/msigdb/genesets.jsp?collection=GO
Pathways/ontologies	MSigDB	Hallmark	http://www.gsea-msigdb.org/gsea/msigdb/genesets.jsp?collection=H
Cell type/tissue of origin	Human Cell Atlas	Genes with cell- or tissue-type specific expression	https://data.humancellatlas.org/origin
Cell type/tissue of origin	Zhang et al and Chou et al	Genes with cell-type specific expression in OA scRNA-Seq	https://www.immport.org/shared/study/SDY998 and https://www.nature.com/articles/s41598-020-67730-y.
Cell type/tissue of origin	Human Protein Atlas	Genes with high protein expression score within each tissue	https://www.proteinatlas.org/humanproteome/celltype
Markers of necrosis and apoptosis	Marshall et al, Tanzer et al, Yang et al	Proteins measured after apoptosis and necrosis, and classical alarmins	https://pubmed.ncbi.nlm.nih.gov/24401845/, https://www.sciencedirect.com/science/article/pii/S2211124719317449 and https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5699517/

Subcellular location	Human Protein Atlas	Location within cell	https://www.proteinatlas.org/humanproteome/cell/organelle
Subcellular location	ExoCarta	Exosome-associated proteins	http://www.exocarta.org/

Table 2: Gene sets to test for enrichment

The first category of gene sets we will consider are canonical pathways and other ontologies. We will take these genesets from MSigDB, and will include KEGG, Gene Ontology and Hallmark gene sets.

The second category will be cell and tissue type of origin gene sets. We will use single-cell RNA-seq data (counts matrices and pre-computed cell type annotations) taken from the Human Cell Atlas to define cell-type-specific gene sets using EWCE (taking the top 10% most specific genes for each cluster). For joint-specific scRNA-Seq data, we will eventually like to rely on the HCA joint atlas data, but as that may not be available in time, we will also run an initial analysis using the single-cell RNA-Seq from OA samples generated by Zhang et al[7] and Chou et al[8]. We will also use pre-defined tissue-of-origin annotations produced by the Human Protein Atlas based on protein arrays.

We will also test gene sets that may indicate the rate and type of cell death, including sets derived from high-throughput proteomic experiments of apoptotic and necrotic cells[9,10], and from curated lists of alarms released during necrosis[11].

Finally, we will generate gene lists based on subcellular location. We will use the pre-computed Human Protein Atlas data to assign organelle locations to individual proteins based on immunohistochemistry, and will test gene sets at the broad level (nucleus vs cytoplasm vs secretary) as well as at the specific organelle level. Finally, we will use gene sets of exosome-associated proteins, to detect whether individual clusters are characterized by exosomal transport.

We will also run a separate analysis using the package *QuaternaryProd*[12] to uncover upstream regulators of protein expression within specific endotypes. Regulatory relationships will be taken from STRINGdb[13]. We will also test for enrichment of heritability for osteoarthritis in the genes present in each cluster using partitioned LD Score regression[14] and summary statistics from the latest OA GWAS meta-analysis[15].

We will visualize these results using bubble charts of significantly differentially expressed gene lists for each gene set category and endotype. We will also visualize the protein

[7] https://www.nature.com/articles/s41590-019-0378-1

[8] https://www.nature.com/articles/s41598-020-67730-y

[9] https://pubmed.ncbi.nlm.nih.gov/24401845/

[10] https://www.sciencedirect.com/science/article/pii/S2211124719317449

[11] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5699517/

[12] https://www.bioconductor.org/packages/release/bioc/html/QuaternaryProd.html

[13] https://string-db.org/cgi/download

[14] https://github.com/bulik/ldsc

[15] https://www.nature.com/articles/s41588-018-0327-1

co-expression network and enriched pathways using *CytoScape*[16] and *igraph*[17], *clusterProfiler*[18] and *enrichplot*[19] packages in R.

Sub-analysis 1.4: Clinical characteristics of endotypes

Questions:

- How do the proteomic clusters correlate with clinical features? Including:
    - Sex
    - Age at sample
    - BMI at sample
    - Smoking history
    - Injury vs OA
    - presence/absence of unacceptable knee pain (PASS)
    - continuous knee pain scores (WOMAC pain subscore for OA, KOOS pain subscore for joint injury).
    - presence/absence radiographic OA (defined as KL grade ≥2/<2)
    - presence/absence of advanced radiographic OA (defined as KL grade ≥3/<3)
    - radiographic severity of OA (KL grade, 0-4)
    - combined pain/radiographic categorization (four categories based on presence/absence of unacceptable pain (PASS) and presence/absence of radiographic OA, ≥2/<2)
- Does this clustering exist in synovial fluid from healthy samples or inflammatory arthritis patients?

Results:

- A table showing associations between each proteomic cluster and each clinical feature: p values on each endotype for each clinical feature.
- A confusion matrix of cluster assignments across repeated samples from the same individuals at different times

Plots:

- Barplots showing the distribution of categorical features across different endotypes
- Violin plots showing the distribution of continuous clinical features across different endotypes.
- PCA/UMAP plots of clustering overlayed with labelled clinical features.

Methods:

We will test for correlation between individual cluster membership and categorical (sex, smoking history, injury/OA, presence/absence of unacceptable pain, presence/absence of radiographic/advanced OA, combined pain/radiographic categorization), continuous (age, pain score, BMI) and categorical (KL grade) demographic and clinical features using logistic regression. For combined presence/absence of unacceptable pain (defined using PASS) and radiographic OA, we will run separate analyses comparing patients who have presence of both to presence of only one (very few patients have absence of both). We will include cohort as a covariate to control for cohort-specific batch effects. In each case, the p-value will be

[16] https://cytoscape.org/

[17] https://igraph.org/r/

[18] https://bioconductor.org/packages/release/bioc/html/clusterProfiler.html

[19] https://bioconductor.org/packages/release/bioc/html/enrichplot.html

calculated using a likelihood ratio test, and the p-value threshold will be determined by Bonferroni correction for all tests taken.

We will also fit a multiple multinomial regression model, predicting endotype category by including all clinical characteristics above in a single model.

We will visualise the correlation between proteomic clusters and clinical features using bar plots/violin plots of variables for each cluster, as well as by overlaying cluster identity and clinical features on a PCA/UMAP plot.

To test whether the endotypes generalize to non-OA and non-joint injury patients, we will assign the inflammatory and healthy control samples to clusters using the nearest centroid classifier using centroids from the baseline samples generated in 1.1. We will also project these samples onto the reduced dimensions (PCA and UMAP) and visualize their positions relative to baseline samples.

Sub-analysis 1.5: Correlation of technical confounders with endotypes

Questions:

- How do the proteomic clusters correlate with technical confounders? Including:
  - Blood staining (measured or imputed)
  - Bimodal dropout signal status
  - Tranche and processing batch
  - Plate, position in plate, run date
  - Sample age, volume and number of freeze-thaw cycles

Results:

- A table showing associations between each proteomic cluster and each technical confounder: p values on each endotype for each confounder.

Plots:

- Barplots showing the distribution of categorical technical confounders across different endotypes.
- Violin plots showing the distribution of continuous technical confounders across different endotypes.
- PCA/UMAP plots of clustering overlayed with labelled technical confounders.

Methods:

The same analysis approach will be used as in Sub-analysis 1.4, with a few modifications:

1. Technical confounders that are expected to differ across cohorts will be adjusted for by including cohort as a covariate in the regression analysis. These are: blood staining, bimodal signal status, processing batch, sample age, sample volume, number of freeze-thaw cycles.
2. Processing batch will be analysed using a mixed model with processing batch as a random effect, using the R package *lme4*.

Secondary analysis: Synovial proteomics of OA clinical features

Secondary analysis questions: for each clinical feature of interest (injury vs OA, pain, structural severity), which proteins correlate with this feature? What pathways and biological functions characterise these protein lists?

Sub-analysis 2.1: Finding SF correlates of clinical features

Questions:

- Which specific proteins correlate with each clinical and demographic feature?
    - Injury vs OA
    - presence/absence of unacceptable knee pain (PASS)
    - continuous knee pain scores (WOMAC pain subscore for OA, KOOS pain subscore for joint injury).
    - presence/absence radiographic OA (defined as KL grade ≥2/<2)
    - presence/absence of advanced radiographic OA (defined as KL grade ≥3/<3)
    - radiographic severity of OA (KL grade, 0-4)
    - combined pain/radiographic categorization (four categories based on presence/absence of unacceptable pain (PASS) and presence/absence of radiographic OA, ≥2/<2)
- Are any of these associations driven by known confounders (BMI and smoking history)?
- Do these proteins fall into correlated sets, or are they independent?

Results:

- Regression coefficients, p-values and standard errors for each protein against each clinical feature (adjusted for age, sex and cohort and adjusted additionally for BMI and smoking history)
- Lists of signature proteins for each of the clinical features
- Lists of co-expressed protein modules that associate with each clinical feature.
- Grouping of signature proteins into co-expressed modules.

Plots:

- Violin plot of expression levels of signature proteins for different clinical features.
- Heatmap showing associations between protein expressions and features.

Methods:

This sub-analysis largely mirrors sub-analysis 1.2. The primary differences are:

- In all regression models, we will test log protein expression as the predictor.
- When testing for protein differences between injury vs OA, we will use logistic regression. When testing for protein differences across the continuous measures (including pain scores) we will analyse it using linear regression, and for ordinal measures (KL grade), we will carry out ordinal regression. These same approaches will be used to test for associations with the eigenproteins generated in sub-analysis 3.1.
- We will check for deviations from additive linearity using diagnostic plots (including one for age), and if deviations are observed we will apply spline regressions using the R package *gam*[20].
- We will include age, sex and cohort in each regression to control for confounding. We will include a secondary analysis conditioning on BMI and smoking history, for samples where this data is available, as well as a secondary analysis conditioning on imputed blood staining.
- We will also carry out robustness analyses conditioning on each of the technical confounders in Sub-analysis 1.5 (each variable included one-by-one as a fixed-effect covariate, with the exception of processing batch which will be included as a random effect).

[20] https://cran.r-project.org/web/packages/gam/

- For the joint injury vs OA analysis, cohort and outcome are perfectly colinear (as OA and injury samples are from different cohorts), and thus we cannot include both in a linear model. For this analysis, we will instead condition only on age and sex, and then carry out a secondary analysis using a random intercept term for cohort in a mixed model using the R package *lme4*[21].
- We will also produce plots of key differentially abundant proteins across injury, OA, inflammatory arthritis and healthy samples (though sample size will be too low to effectively test for differences in the latter two categories).

Sub-analysis 2.2: Bioinformatic characterisation of clinical features

Questions:

What are the functional/biological characteristics of these protein correlates?

- Which canonical pathways or functions are enriched in the clinical-feature-associated protein lists?
- Are individual clinical features associated with proteins characteristic of a particular set of cell types?
- Do the proteins that associate with each clinical feature localize to a particular part of the cell?
- Which upstream modulators can explain differences in protein levels between the clinical groups?
- Are the proteins associated with the clinical features enriched for heritability in genome-wide association studies (GWAS) of OA?

Outputs and Methods are the same as for sub-analysis 1.4.

Network analysis: Co-expression and regulation of synovial fluid proteins in OA and injury

Network analysis questions: Are there groups of co-expressed proteins in synovial fluid? What are the biological functions of these co-expressed proteins? Do these correlation patterns differ between different clinical groups? Can regulatory relationships and physical interactions account for this co-expression?

Sub-analysis 3.1: Co-expression analysis

Questions:

- How many groups (modules) of co-expressed proteins are there in synovial fluid?
- What are the biological functions of the proteins in each module?
- Are co-expression patterns conserved between OA and injury, and between low and high radiographic severity?

Results:

- An adjacency matrix for each gene co-expression network (OA, injury, KL < 2, KL >=2), with edges given by the Spearman's rank correlation.
- Lists of proteins included in each module selected by WGCNA.
- Enrichment p-values for each gene list in Table 2 for each module.
- Network similarity metrics and permuted Z-scores between pairs of modules across different clinical groups (OA vs injury, KL >= 2 vs KL <2).

[21] https://cran.r-project.org/web/packages/nlme/index.html

Plots:

- Network plots, coloured by module and submodule.
- Pathway enrichment plots for each module
- Module comparison plots for OA vs injury, and for KL >= 2 vs. KL < 2.

Methods:

We will use the R package *GWENA*[22] to carry out co-expression analyses, following the steps recommended in the GWENA vignette[23]. In brief: we will generate unsigned correlation networks using Spearman's correlation for baseline OA and joint injury samples. These networks will be split into modules by hierarchical clustering, with the number of modules determined using the 'hybrid' approach in the R package *dynamicTreeCut*, with parameter value deepSplit = 2, followed by merging of modules with correlated eigenproteins (correlation coefficient > 0.75). To test for enrichment of function pathways in each module, we will test for overrepresentation of genes in the gene lists in Table 2 in each of the modules using g:Profiler[24], with associations declared significant if they meet p < 0.05 after multiple testing correlation using the g:SCS algorithm. To test whether patterns of co-expression are shared across injury and OA samples, we will use GWENA's comparison-by-permutation approach (testing both injury modules in OA and OA modules in injury). We will also generate modules separately for high and low radiographic severity OA (KL >= 3 and KL < 2), and test whether modules are preserved between these patient groups.

We will also test for association between the eigenprotein for each module with discovered endotypes and with clinical characteristics, as described in Sub-analysis 1.2 and 2.2.

Sub-analysis 3.2: Investigating drivers of synovial protein co-expression

Questions:

- What are the most central ("hub") proteins in the inferred co-expression networks? Do these differ between injury and OA, between different radiographic OA severities?
- Do highly coexpressed proteins tend to physically interact?
- Are co-expression effects driven by direct regulatory effects? If so, do these regulatory relationships differ in different clinical groups?

Results:

- Weighted centrality measures (gene connection significance values) for each protein in the co-expression network for each group (OA, injury, KL >= 2, KL < 2).
- Unweighted centrality measures (degree and betweenness) for each SomaScan protein in the protein-protein physical interaction network.
- Network similarity metrics and permuted Z-scores between co-expression and physical interaction networks.
- Directed adjacency matrix of gene regulatory network (GRN) for each group, and similarity metrics for GRNs between clinical groups.

[22] https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-021-04179-4
[23] https://www.bioconductor.org/packages/release/bioc/vignettes/GWENA/inst/doc/GWENA_guide.html
[24] https://cran.r-project.org/web/packages/gProfileR/index.html

Plots:

- Network plots of co-expression data with hub proteins highlighted
- Combined network plots of physical and co-expression data, with hub proteins for both highlighted
- Network comparison plots for co-expression and physical interaction networks
- Box plots of physical PPI centrality statistics for hub and non-hub genes in the co-expression network
- Directed network plots of GRNs
- Network comparison plots of GRNs between different groups

Methods:

To identify hub proteins (i.e. proteins that have high centrality in the network), we will use the gene connection significance values implemented in the R package *dhga*[25] (defining hub proteins as those that have a connection p < 1e-5). We will use the same package to compare overlap in hub proteins across different conditions.

To find out whether co-expression relationships mirror physical protein interaction, we will test for similarity between co-expression network and STRING physical PPI network using the *NetworkDistance*[26] package in R. We will also test whether hub proteins in the co-expression network have significantly higher degree or betweenness statistics in the physical PPI network than non-hub proteins.

We will infer directed gene regulatory networks (GRNs) from abundance data using the *GENIE3*[27] R package. We will compare inferred networks across clinical groups (OA vs injury, KL >= 2 vs. KL < 2).

Networks will be visualized using igraph, GWENA and Cytoscape, where appropriate.

[25] https://cran.r-project.org/web/packages/dhga/

[26] https://cran.rstudio.com/web/packages/NetworkDistance/index.html

[27] https://bioconductor.org/packages/release/bioc/html/GENIE3.html

Appendix 1: Data releases associated with this data analysis plan

The discovery analysis plan will make use of the datasets described below.

discovery_somascan_1: Proteomic quantification, normalization, batch correction and QC

This dataset includes SomaScan protein quantification of 7596 somamers for 6627 proteins for 1045 synovial fluid samples. Protein intensity data is normalized, and then the log intensity batch corrected by plate using the ComBat function in the *sva* R package[28]. Files released include raw, normalized and batch-corrected data. Throughout the discovery analysis plan above, the ComBat batch-corrected data should be used as the primary analysis, with the normalized-but not-batch-corrected data used for secondary robustness tests.

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**	**Primary (batch-corrected)**	**Secondary (non-batch corrected)**
NONHUMAN	Non-human proteins	Non-human or control proteins	Proteins	307	307
OA_REPO	Reproducibility in OA pool	Predicted R2 < 0.5	Proteins	485	485
INJ_REPO	Reproducibility in injury pool	Predicted R2 < 0.5	Proteins	252	252
VOLUME_CONFOUND	Associated with sample volume	ANOVA p < 0.05/7289 (conditional on cohort)	Proteins	488	413
FREEZETHAW_CONFOUND	Associated with number of freeze-thaw cycles	ANOVA p < 0.05/7289 (conditional on cohort)	Proteins	83	50
SAMPLEAGE_CONFOUND	Associated with sample age	ANOVA p < 0.05/7289 (conditional on cohort)	Proteins	263	730
BIMODAL_CONFOUND	Associated with bimodal signal	ANOVA p < 0.05/7289	Proteins	96	1947
SOMASCAN_FAIL	SomaLogic inhouse QC	Hybridization Scale Factor > 2.5	Samples	2	2
LOD_SAMPLE	Limit of detection	25% of proteins below/above limit of detection	Samples	12	12
TOTPROT_OUTLIER	Total protein outliers	>5SDs from mean	Samples	9	9
PCA_OUTLIER	PCA outliers	>5SD from combined centre on top PCs	Samples	15	15
	Total remaining		**Proteins** **Samples**	**5944/7596** **1030/1045**	**4734/7596** **1030/1045**

[28] https://rdrr.io/bioc/sva/man/ComBat.html

discovery_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 (tranche1 vs tranche2)	{1 = tranche 1, 2 = tranche 2}
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 sample was hyaluronidase treated by KIR	Text

date		(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

Note on reading the SIN: The SIN is in the format:

STEP[NNNN]-V[N]-F[-R/L]-HT[N][-SP/UN]

STEP[NNNN] is a unique anonymous patient identification number (PIN), V[N] gives the visit number where the sample was collected (V1 for baseline visit, V[2+] for later visits), F gives the sample type (all SomaScan samples are F, for synovial fluid), [-R/L] gives the right or left hand side (present for 2x samples in tranche 1/2 (STEP1433 & STEP2001) and all samples in tranche 3), HT[N] gives the sample processing number (HT1 is the first processing of this sample, and so on), and [SP/UN] demarks spun or unspun samples (only present for spun/unspun pairs, otherwise all samples in this dataset can be assumed to have been spun).

discovery_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	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).	{0 = No, 1 = Yes}

discovery_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.}

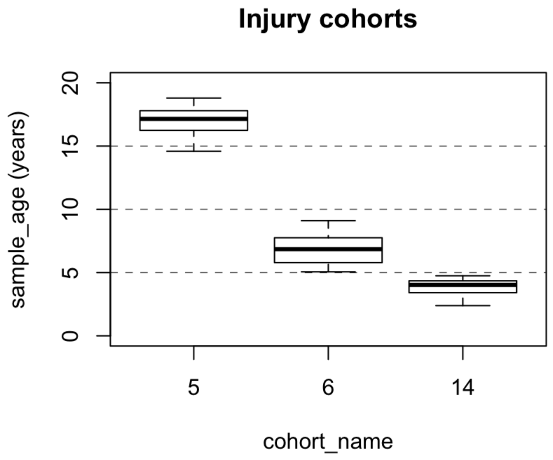
harm_pain_prom	The specific patient reported outcome measure used to derive harm_knee_pain.	{1 = KOOS, 2 = WOMAC, 3 = Knee specific VAS/NRS, 4 = PainDETECT VAS, NA=missing}
koos_pain	KOOS pain subscore (calculated from full KOOS questionnaire results, or from combined subscore provided by cohort collectors, 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

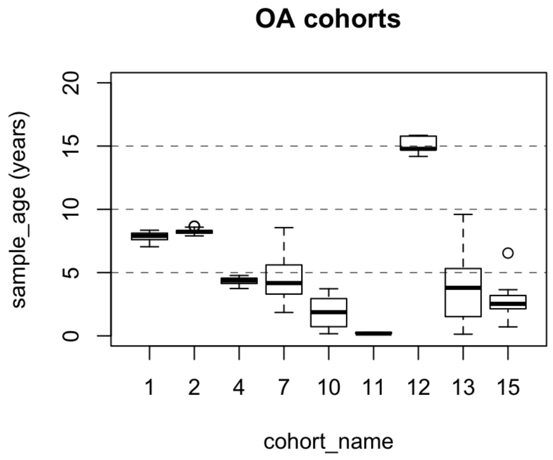
Appendix 2: Descriptive statistics for the clinical data

The tables and graphs below give the missingness statistics and the distribution across cohorts for the released phenotype data. Note that this data is specifically for baseline samples.

QA variables

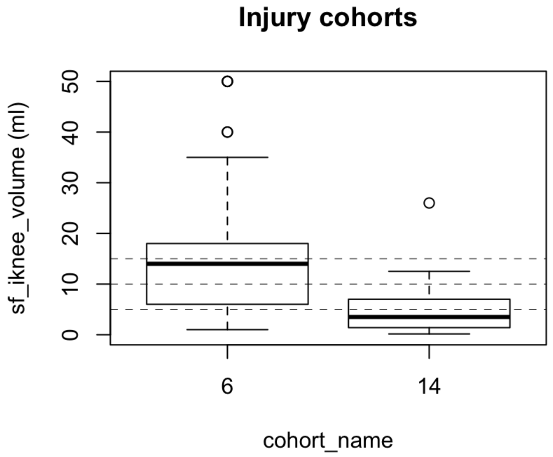
Age of sample (one sample with missing data)

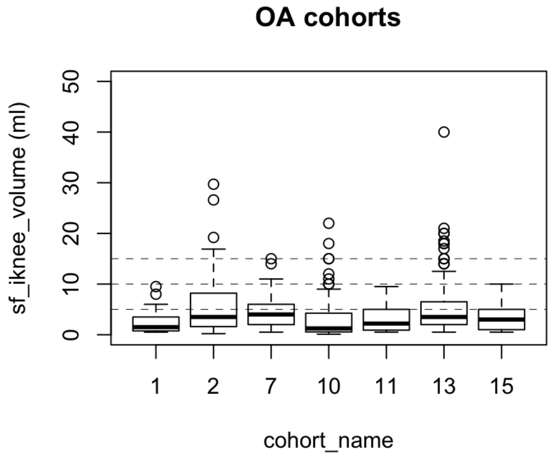




Volume of sample

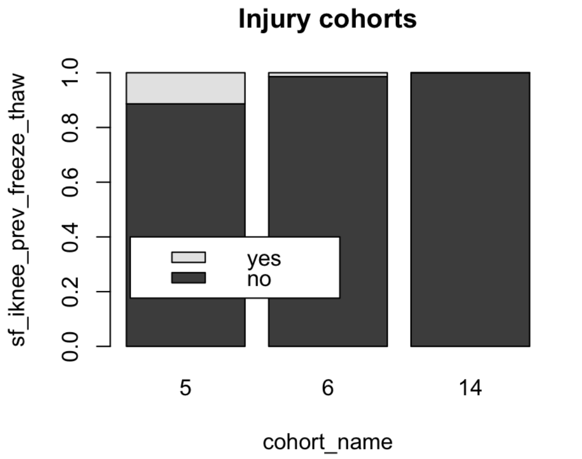
sf_iknee_qc_group	sf_iknee_volume not missing	sf_iknee_volume missing
0 (OA)	671	48
1 (Injury)	172	46
2 (Healthy)	1	36
3 (Inflammatory)	0	5

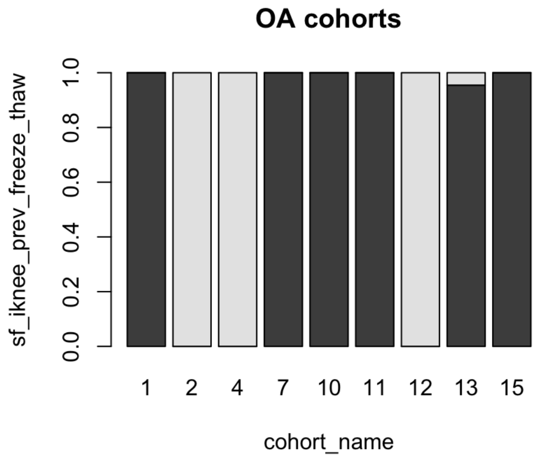




previous freeze-thaw

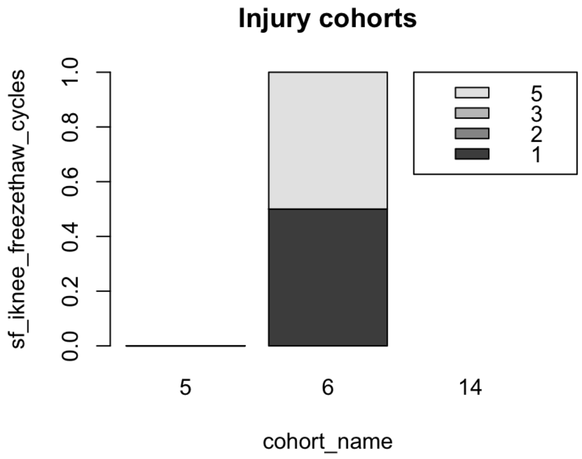
sf_iknee_qc_group	0 (No)	1 (Yes)	2 (Not known)
0 (OA)	642	76	1
1 (Injury)	211	7	0
2 (Healthy)	0	33	4
3 (Inflammatory)	0	0	5

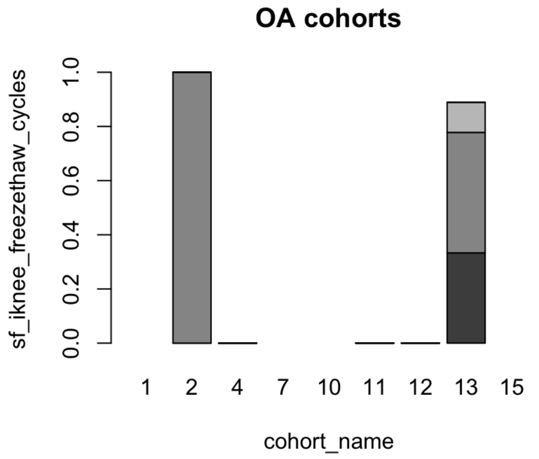




Number of freeze-thaw cycles (for those that underwent at least one freeze-thaw)

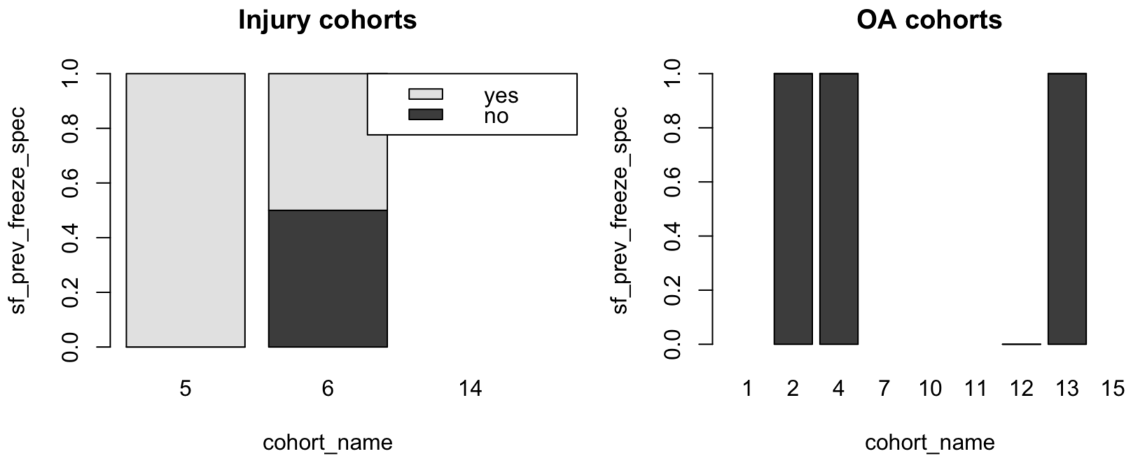
sf_iknee_qc_group	1	2	3	5	Missing
0 (OA)	4	49	1	0	22
1 (Injury)	1	0	0	1	5
2 (Healthy)	0	2	0	0	31





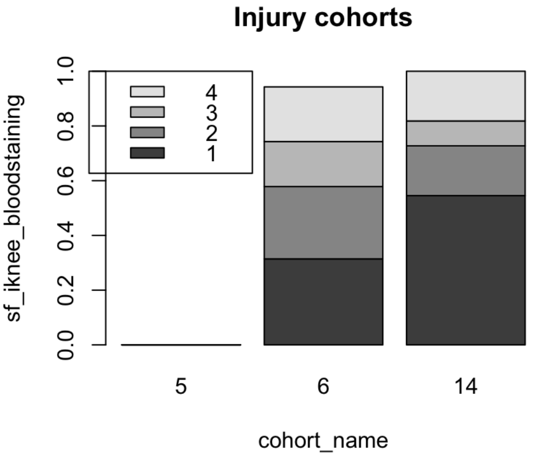
Five or more freeze-thaws (of those that underwent at latest on freeze-thaw)

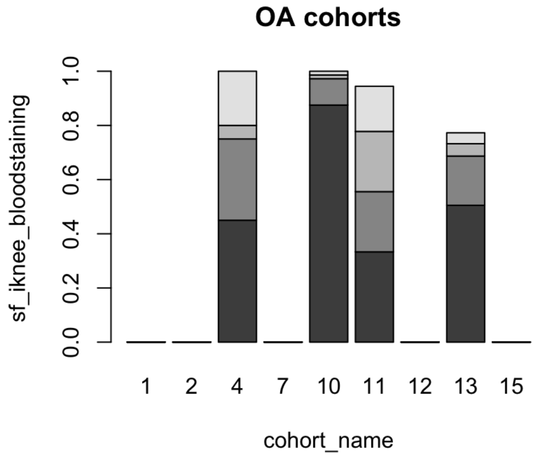
sf_iknee_qc_group	0 (<5 freeze-thaws)	1 (>=5 freeze-thaws)	Missing
0 (OA)	74	0	31
1 (Injury)	1	6	0
2 (Healthy)	2	0	1



Blood staining

sf_iknee_qc_group	sf_iknee_bloodstaining present	sf_iknee_bloodstaining missing
0 (OA)	265	454
1 (Injury)	165	53
2 (Healthy)	4	33
3 (Inflammatory)	5	0

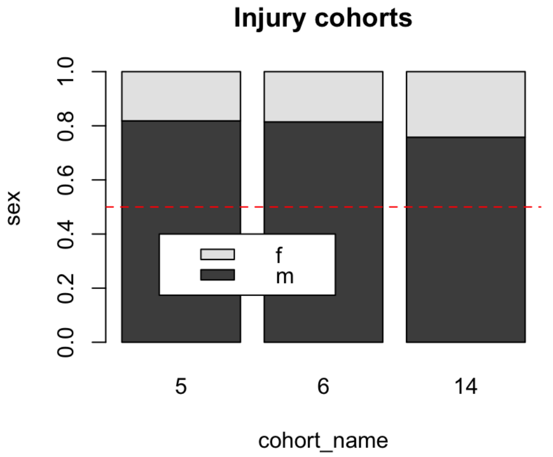


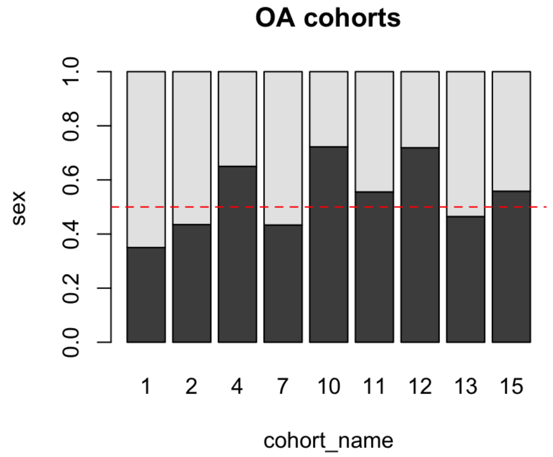


Demographic variables

sex

sf_iknee_qc_group	sex == "m" (male)	sex == "f" (female)	sex == "" (missing)
0 (OA)	348	371	0
1 (Injury)	176	42	0
2 (Healthy)	25	11	1
3 (Inflammatory)	2	3	0

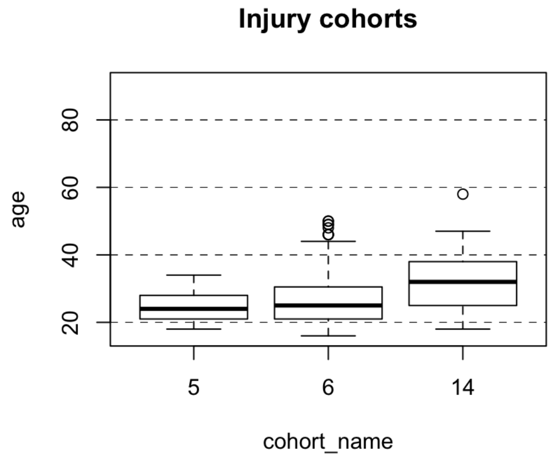


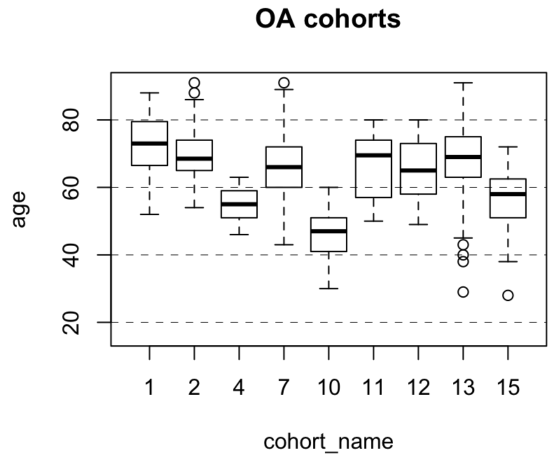


age

sf_iknee_qc_group	age present	age missing
0 (OA)	718	1

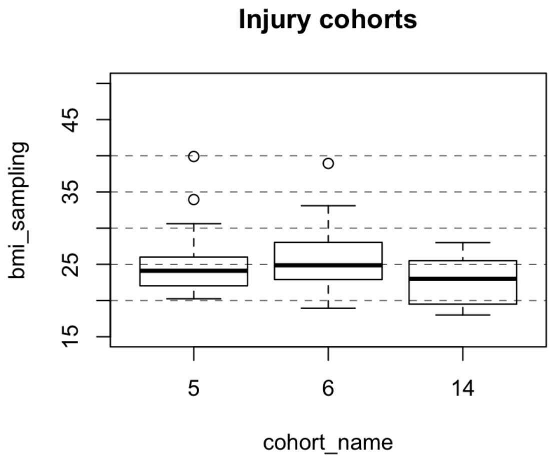
1 (Injury)	218	0
2 (Healthy)	36	1
3 (Inflammatory)	5	0

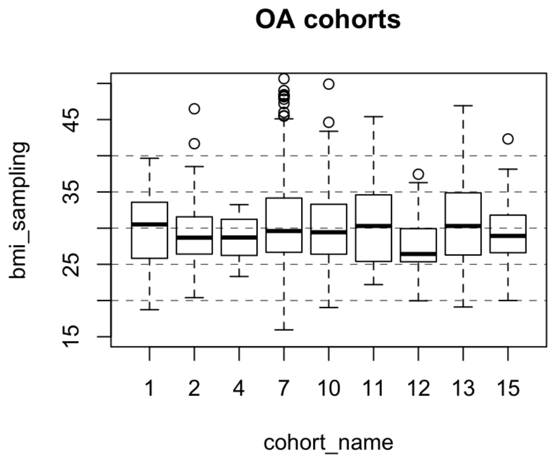




bmi

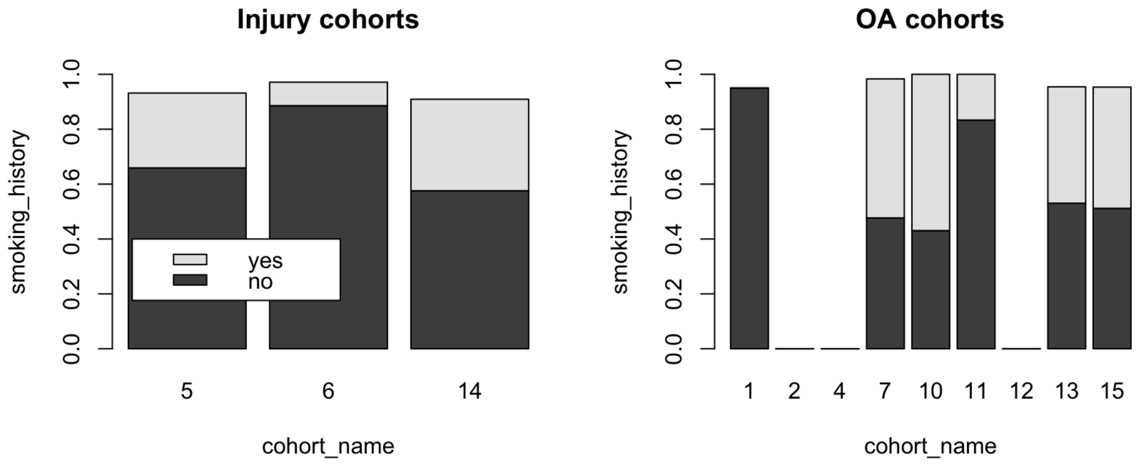
sf_iknee_qc_group	bmi_sampling present	bmi_sampling missing
0 (OA)	705	14
1 (Injury)	190	28
2 (Healthy)	31	6
3 (Inflammatory)	0	5





smoking history

sf_iknee_qc_group	0 (No)	1 (Yes)	2 (Not known)
0 (OA)	334	298	87
1 (Injury)	172	35	11
2 (Healthy)	0	0	37
3 (Inflammatory)	0	0	5

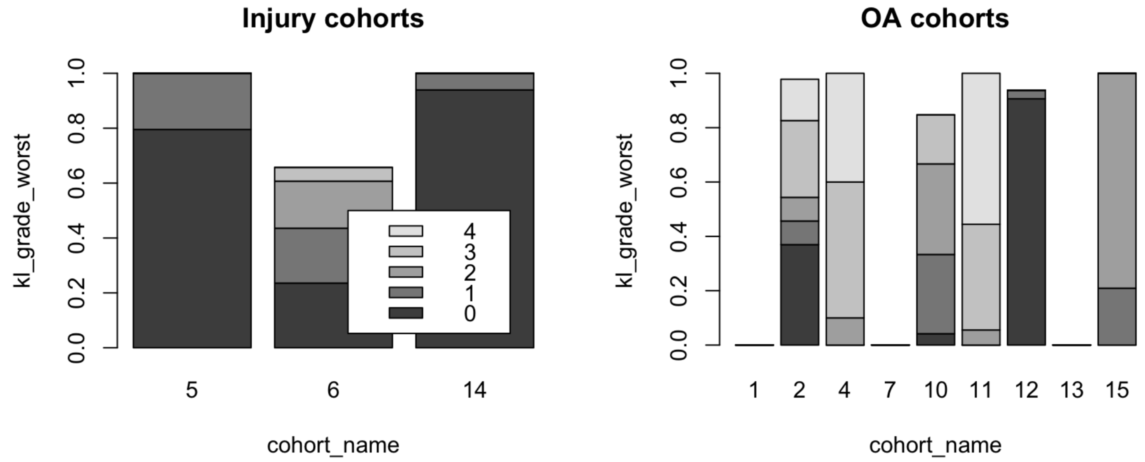


Radiographic variables

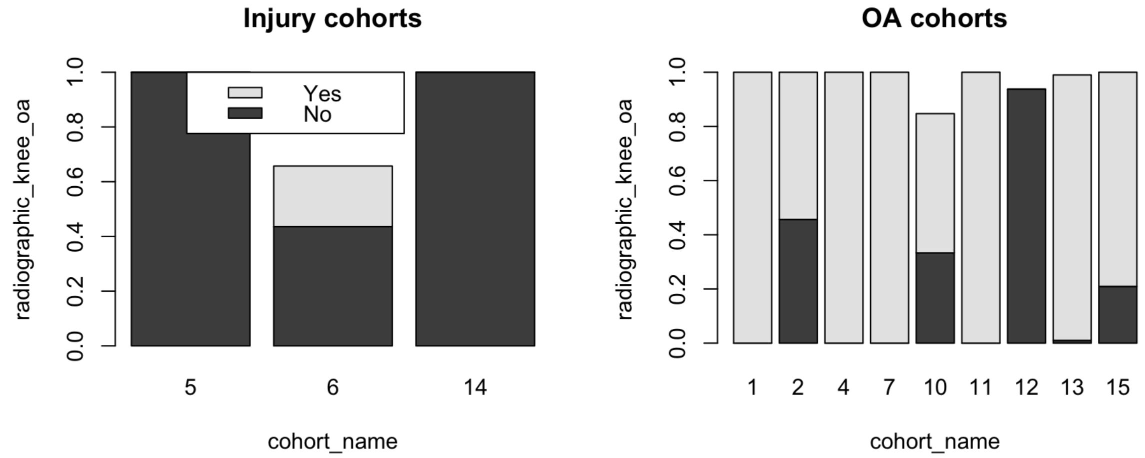
KL grade fields

sf_iknee_qc_group	kl_grade_worst known and non-missing	radiographic_knee_oa known and non-missing	kl_grade_advanced known and non-missing	all KL fields Not Known or missing
0 (OA)	188	706	716	3
1 (Injury)	169	170	218	0
2 (Healthy)	29	30	30	7
3 (Inflammatory)	0	0	0	5

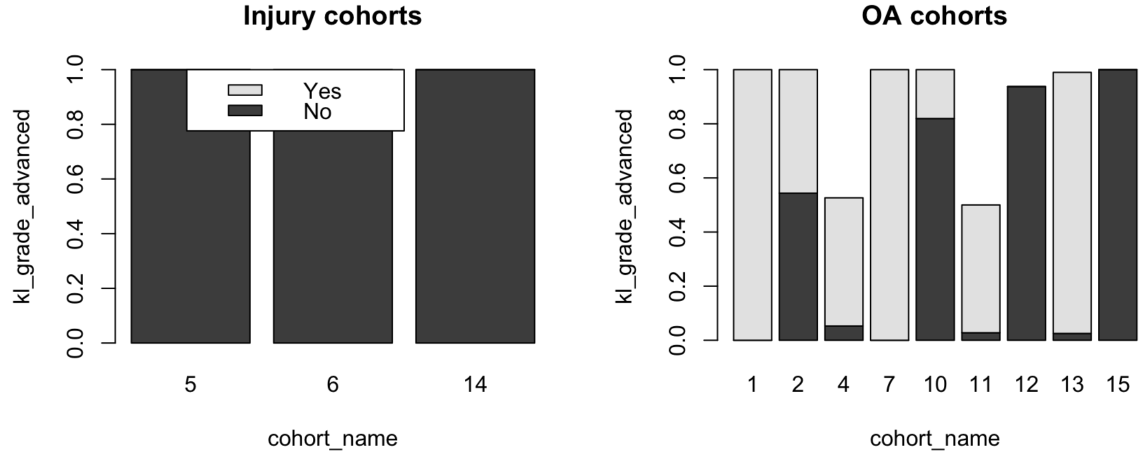
Raw KL grades



Radiographic OA flag



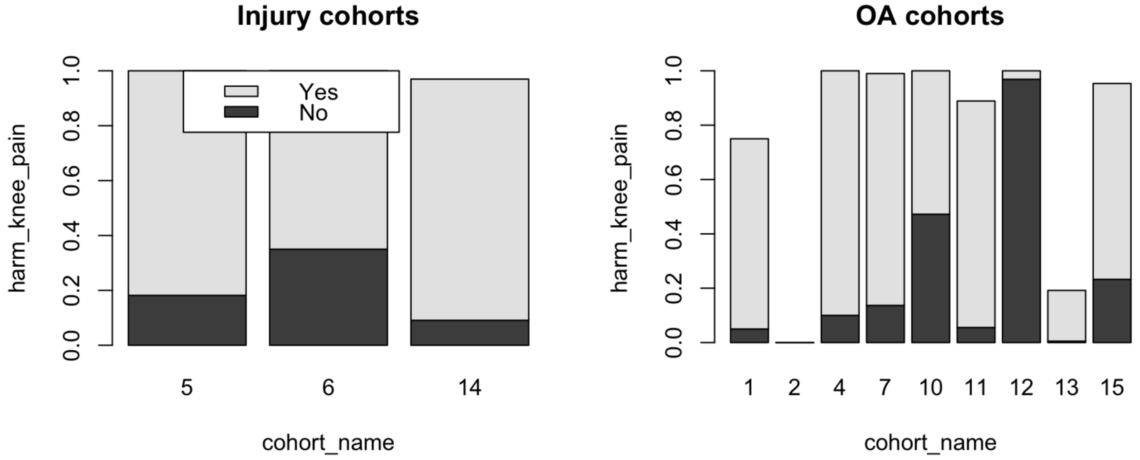
Advanced radiographic OA flag



Pain variables

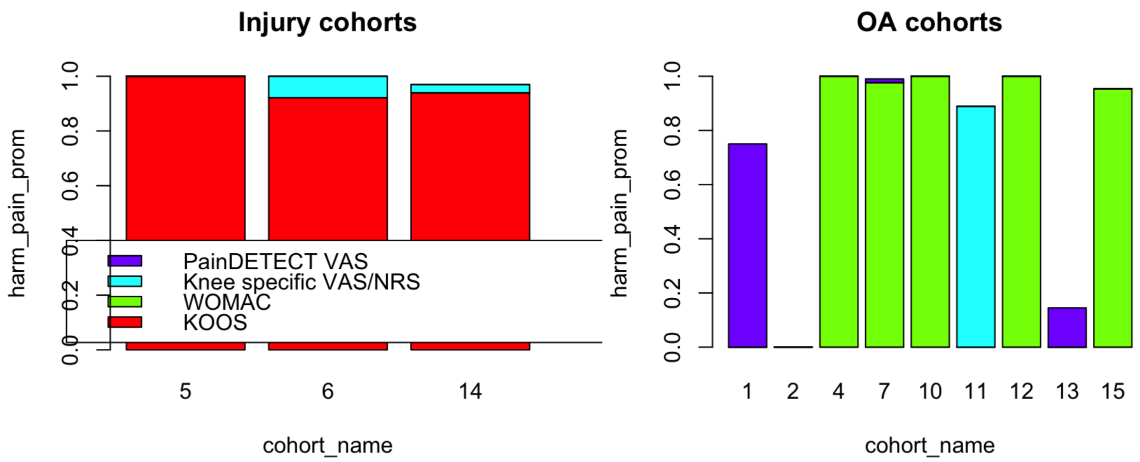
Unacceptable (above PASS) pain

sf_iknee_qc_group	harm_knee_pain == 0 (No)	harm_knee_pain == 1 (Yes)	harm_knee_pain == 2 or missing (Unknown)
0 (OA)	91	409	219
1 (Injury)	60	156	2
2 (Healthy)	30	1	6
3 (Inflammatory)	0	0	5



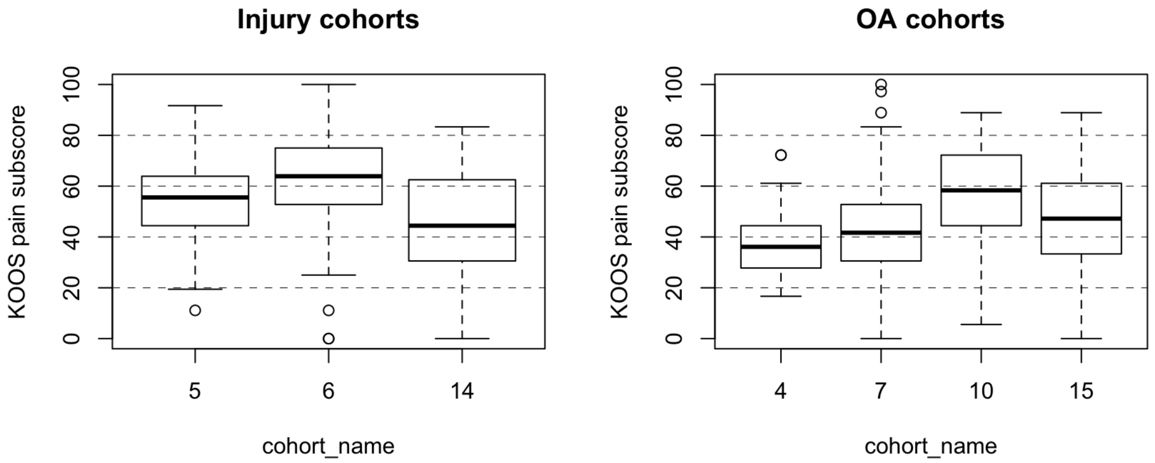
PROM used to calculate thresholded pain

sf_iknee_qc_group	KOOS	WOMAC	Knee specific VAS/NRS	PainDETECT VAS	Missing
0 (OA)	0	427	16	57	219
1 (Injury)	204	0	12	0	2
2 (Healthy)	0	31	0	0	6
3 (Inflammatory)	0	0	0	0	5



KOOS pain subscore data

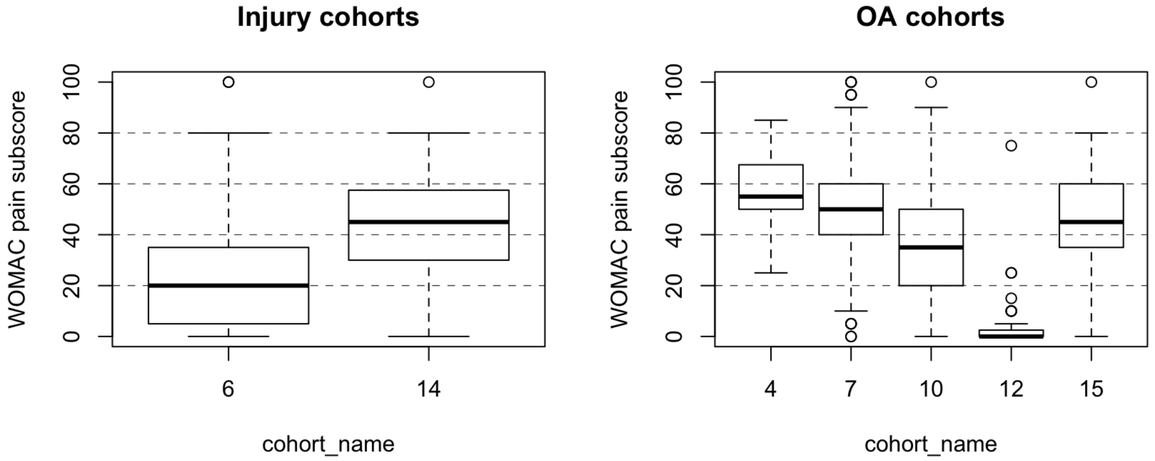
sf_iknee_qc_group	koos_pain nonmissing	koos_pain nonmissing
0 (OA)	419	300
1 (Injury)	204	14
2 (Healthy)	0	37
3 (Inflammatory)	0	5



WOMAC pain subscore data

sf_iknee_qc_group	womac_pain nonmissing	womac_pain missing

0 (OA)	427	292
1 (Injury)	161	57
2 (Healthy)	31	6
3 (Inflammatory)	0	5



Appendix 3: Power calculations

In the interests of brevity, we will not present a detailed power calculation for every sub-analysis detailed above. Instead, the figures below show power assessments for a few key analyses, to aid interpretation of analysis results.

Figure A3.1 shows the power that our analysis approach will have to detect significant (f_k < 0.85) clustering of OA samples, assuming a simple clustering structure across 10 principal components, with varying strength of clustering and true number of underlying clusters. We have provided reduced dimensional representations of three example clustered datasets that represent scenarios where our analysis would have high power (>80%) to detect clustering if there were two, three or four underlying clusters. If two clusters exist in the data, we will have high power to detect them even if the clustering is relatively subtle, but as the number of clusters grows we require stronger overall clustering to be well powered.

Figure A3.2 demonstrates the power for the more stringent task of detecting significant clustering and assigning the correct number of clusters. For two true clusters, this is similar to the power for detecting significant clustering alone, but for three or four true clusters the power is limited even for a large number of clusters. The conclusion of this analysis is that the number of assigned clusters should be considered a lower limit rather than a good estimate of the true number.

Figure A3.3 shows the power to detect significant correlations between protein concentration and a continuous clinical variable (such as the continuous pain score) in the OA and injury samples. This analysis stands here as a proxy for, and upper limit on, the power for associations between proteins and clinical variables in general. The OA analysis shows >80% power to detect associations with a correlation coefficient >0.19 (i.e. where the protein can account for >3.6% of variance in the continuous clinical variable). For the injury analysis, this correlation coefficient value is >0.34 (i.e. accounting for >11.9% of variance).

Methods

The power to detect significant clustering, and assign the correct number of clusters, was assessed using simulation. To simulate the reduced dimensional space, we used a 10-dimensional Gaussian mixture model with cluster frequencies sampled from dirichlet(alpha = 10), and covariance matrices for each cluster sampled from InverseWishart(20,I_{10}). To simulate a varying degree of sparseness and strength of clustering, we sampled mean vectors with the first 3, 4 or 5 values sampled from N(0,sigma^2), where sigma^2 was varied in the range 0-5, and the remaining values of the mean vector set to zero. The data was then clustered, and the number of clusters determined, as described in Sub-analysis 1.1.

Power to detect correlation between a protein and a continuous clinical variable was calculated using the *pwr.r.test* function in the package *pwr*, using an alpha value of 0.05/5000.

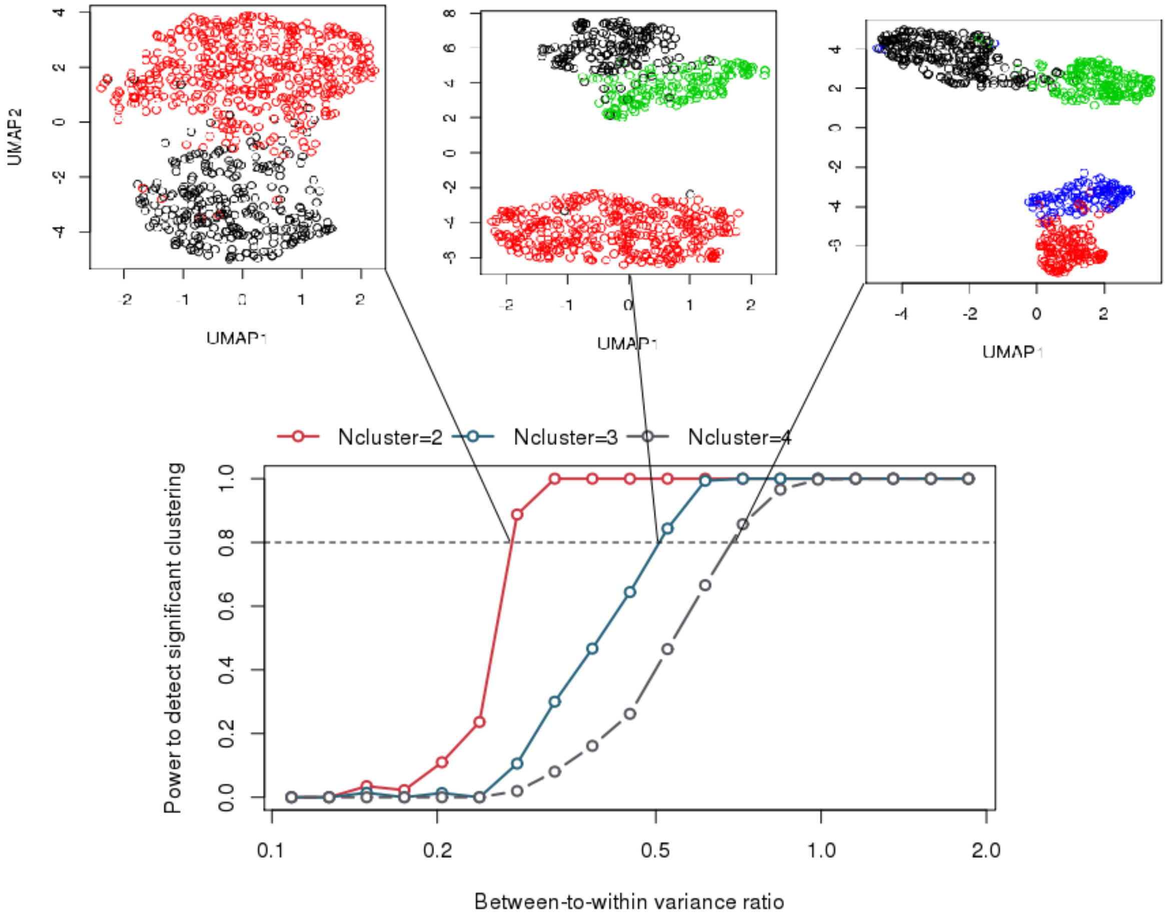


Figure A3.1: Power to discover significant clustering using k-means clustering and the f_k statistic (measured by the proportion with f_k < 0.85), as a function of the number of true clusters and the strength of the clustering, in the OA samples (assuming N=754 samples). UMAPs are shown for example simulated clusters with between-to-within variance ratios that produce a power of 80%.

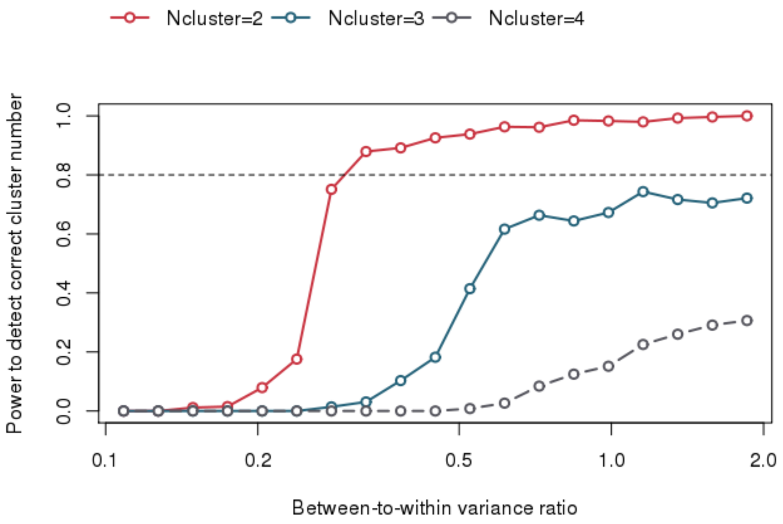


Figure A3.2: Power to correctly determine the true number of clusters using the NbClust voting approach,, as a function of the number of true clusters and the strength of the clustering, in the OA samples (assuming N=754 samples).

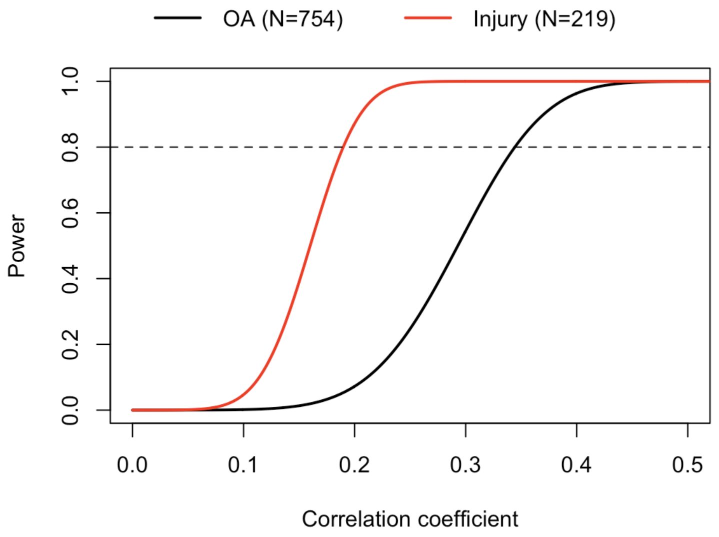


Figure A3.3: Power to detect significant associations between protein quantifications and a continuous clinical trait, after correcting for multiple testing ($p < 0.05/5000$), for OA and injury samples.

Appendix 4: Lay summary of the analysis plan

Osteoarthritis (OA) affects millions of people worldwide, yet we have limited understanding of what causes it and how to select the right drugs to treat it. In addition, OA is highly variable between individuals: some will stay stable for many years with modest pain and disability, yet others will progress over time and require surgical joint replacement. Joint injury also predisposes to the development of OA even in younger individuals, but this risk is also unpredictable. Being able to predict development or progression of disease and identify distinct or shared molecular causes of disease is vital for developing and testing new treatments.

STEpUP OA is a large international effort to attempt to answer some of these questions. We have assembled a group of doctors, scientists and individuals with OA (or at risk of OA) across a number of universities, pharmaceutical companies and hospitals to design this study. We will analyse the knee fluid (obtained by needle and syringe) of nearly 2000 individuals who have a diagnosis of OA or who have recently had an acute knee injury. These samples have already been collected and are stored. We will use cutting edge methods to measure over 5000 different protein molecules in each of the fluid samples and study these alongside patient reported measures such as pain and disability.

We are now starting the first major analysis (which we have called the Discovery Analysis), on the first 1079 samples. Using advanced statistical methods we will be able to address a number of key questions. We will use the signature of proteins in the knee joint fluid to ask whether OA is a single disease at the protein level, or whether there are multiple different types of OA that can be identified. We will seek common pathways which are suggested by specific proteins to identify possible novel causes of disease. We will study the relationship between different protein levels and pain to uncover new molecules that could be targeted for treatment or which could represent an objective marker (biomarker) of pain for future clinical trials.

Finally we will create a rich data set that can be used by the broader OA community to enhance OA research internationally and facilitate the development of new drugs for those with or at risk of OA.

Appendix 4: Sub-Analysis Plan 1: Outcome-Guided Clustering

Author: Laura Bondi and Brian Tom
Date: 05/10/21

Questions to be addressed:

- Using the baseline samples, can we identify subpopulations of patients homogeneous for protein marker profiles such that clusters may have clinical meaning (outcome-guided clustering)?
  - Are there differences in the molecular endotypes that we find in the disease patient groups with different radiographic grades (coded by the variables radiographic_knee_oa, kl_grade_advanced and kl_grade_worst)?
  - How are the proteomic data clustered in knee osteoarthritis (OA) and in acute knee injury?
- Are these data well described by multiple clusters? How many? This is of interest when considering only OA patients (low and advanced grade) but also when including all OA and acute knee injury patients.
  - Do "model-based" (outcome-guided) clustering and unsupervised clustering give similar results in this context? That is, does the clustering remain if all patients are considered together agnostic to the clinical differentiation of patients?
  - Which protein markers have important roles in identifying subgroups of patients with similar radiographic grade or other clinical outcome/phenotypes, such as perceived pain? Pain is coded by the variables harm_knee_pain (binary harmonised knee pain variable), womac_pain (PROM for OA, continuous variable) and koos_pain (PROM for injury, continuous variable).

Methods:

We will carry out Bayesian profile regression (model-based outcome-guided clustering approach) to identify clusters of protein marker profiles that are associated with the clinically relevant outcomes, such as disease radiographic grade (low vs advanced, "disease" group (OA vs acute injury) or pain severity phenotype. This clustering methodology can handle a large number of possibly inter-related explanatory variables and uses the information in both these explanatory variables (i.e. protein markers) and the outcome to produce model-based clustering structures, where the uncertainty associated with these clustering structures and the number of clusters is reflected. Convergence of the MCMC algorithm will be investigated by checking agreement between the "representative" clustering structures for multiple independent chains and a final representative cluster will be determined.

We will explore variable selection and dimension reduction approaches for determining the relevant variables or (combination of variables) that inform the clustering structure. Two-staged and joint strategies will be investigated to determine the most appropriate / informative way of identifying the relevant variables that lead to clinically meaningful clustering.

More specifically, in a two-staged approach we can first screen for protein markers that are differentially expressed between groups of patients defined by the clinical outcome (adjusting the p-values to account for multiple comparison by say using false discovery rate (FDR) or Bonferroni's correction) to be used in the second stage (i.e. application of Bayesian profile regression). Alternatively we can project the space of protein markers into a low-dimensional space (e.g. using PCR or PLS) and use a reduced set of derived components as the new variables to be taken forward to the second stage of Bayesian profile regression. In the joint approach, we simultaneously perform the clustering and the variable selection within Bayesian profile regression.

We will compare results from Bayesian profile regression to the main unsupervised clustering approach, k-means (with dimension reduction), adopted in the SAP and to a sparse k-means clustering approach.

Results:

- Table of descriptive statistics of the sample of patients used (before clustering), separately for OA and acute knee injury patients;
- Cluster assignments for each patient under the best partition and uncertainty associated with such partition;
- Measures of variable importance for the clustering structure;
- Table of descriptive statistics characterising the relevant proteins and outcome profiles of each cluster

Plots:

- Posterior similarity matrix for the consensus across multiple MCMC chains from the Bayesian profile regression analysis;
- Cluster sizes for the final representative consensus clustering;
- Protein and outcome profiles of each clusters from the final representative clustering.