Appendix D


1 AGE AND GENDER-BASED CLUSTER BIAS

1.1 Age distributions

In Appendix B, Figure B5, the age distributions for the three clusters discussed in the paper was compared, showing that there was no statistical difference between clusters in terms of age. As complementary information, the age-vs-gender distributions was investigated, which showed no statistical difference between the male and female sub-cohorts in terms of age (Figure D1).

1.2 Gender-related clustering bias

The clustering results discussed in Section 5.1 were further investigated with respect to gender.

In Figure D2, the markers-vs-gender distributions showed that males had lower concentrations of: S_ARGS, S_C10C, S_COLL2_1, S_COLL2_1NO2, S_CTXI, S_NMID, U_CTXII, and U_CTXI_ALPHA. Males also had lower biomarker values of the S_COLL2_1NO2, U_CTXII, and U_CTXI_ALPHA FNIH/OAI markers (Figure D3). Therefore, the clusters discussed in this paper were expected to present, to some extent, a bias due to the different distribution across gender. The presence of this bias was confirmed by the statistical test for the feature "Sex" provided in Appendix B, Figure B5, which showed that the genders were not distributed evenly across clusters: the low tissue turnover cluster (green) contained approximately 40% males and 60% females; while other clusters contained substantially fewer males than females (structural damage: <10% males, >80% females; systemic inflammation: approximately 20% males and 80% females. The 67 males in the IMI-APPROACH cohort were distributed as follows: 37 were in the low tissue turnover cluster, 10 were in the structural damage cluster and 20 were in the systemic inflammation cluster. In the FNIH/OAI cohort, we found similar results (Figure D4).

Due to this unequal gender distribution, conducting a gender-specific clustering would be appropriate to further investigate how male and female sub-cohorts are stratified by our algorithm. However, due to the low number of males (n=67) in the IMI-APPROACH cohort this investigation was limited, i.e., a gender-based clustering would not be as strong as if performed on the entire cohort. The female sub-cohort was substantially larger than the male sub-cohort, and therefore it most likely was driving the clustering process on the entire cohort.
Despite the above-mentioned limitations, insights for the gender-based clustering results for the IMI-APPROACH male and female sub-cohorts were provided. The proposed pipeline were run for male and female patients separately (excluding the missing data imputation and non-fasting patient correction, which were conducted on the whole population). Figure D5 shows the radar plots obtained from the female and male sub-cohorts. The clustering profiles obtained by the female sub-cohort were indeed similar to those obtained by the whole population. Conversely, the male sub-cohort provided a substantially different stratification compared to what was seen for the females. In particular, male profiles were not discriminable anymore, in terms of S_ARGS, S_COMP, S_PRO_C2, S_C10, S_HA, S_COLL2_1NO2, which was the case in both the female sub-cohort and the whole cohort. Nevertheless, the main fingerprints of a low tissue turnover, systemic inflammation and structural damage related clusters were still present yet weakened. In this sense, despite the common traits of male and female clusters, we believe that the three clusters’ definitions should be different if male and female patients were considered separately.

In conclusion, while more research should be conducted on more abundant cohorts to fully evaluate gender bias in clustering analysis of OA-related biochemical markers, our analysis suggested that gender played a relevant role in driving clustering results. Therefore, we believe it is advisable for future studies to consider male and female patients separately and possibly draw conclusions that are gender-based, if sample sizes are large enough.

**Figure D1.** Age-vs-gender distributions for the IMI-APPROACH patients. There is no statistical difference between the male/female cohorts (Mann-Whitney Test, p=0.2854).
Figure D2. Markers-vs-gender distributions for IMI-APPROACH patients. The male-vs-female statistical difference is assessed via Mann-Whitney tests with the Benjamini-Hochberg correction for multiple comparisons (alpha=0.05). P-values are provided on top of each graph only when the null hypothesis is rejected, i.e. the two distributions are statistically different.
Figure D3. Markers-vs-gender distributions for FNIH/OAI patients, only those markers that are in common with the IMI-APPROACH cohort. The male-vs-female statistical difference is assessed via Mann-Whitney tests with the Benjamini-Hochberg correction for multiple comparisons (alpha=0.05). P-values are provided on top of each graph only when the null hypothesis is rejected, i.e. the two distributions are statistically different.
Figure D4. Male (1) and female (2) distribution within clusters found in the FNIH/OAI cohort, as discussed in Appendix C. Pair-wise p-values are reported on top of each statistically different pair (Chi-square test). The distributions are similar to those found in the IMI-APPROACH cohort, Appendix B, Figure B5.
Figure D5. Radar plots comparing clusters found in the female and male sub-cohort in IMI-APPROACH. Bold dots represent median values that are statistically different to the other two values into the same axis. For further information how to interpret these plots, refer to Figure 5.