Data Harmonisation across Databases for Machine Learning Model Development to Predict Knee Osteoarthritis Progression

AUTHORS: S. Danso[1], P. Widera[1], P. M. Welsing[2], S. Peelen[3], A. Tielmann[4], F. Berenbaum[5], F. J. Blanco[6], I. K. Haugen[7], L. Hussaarts[3], M. Kloppenburg[8], F. Lafeber[2], J. Larkin[9], M. C. Levesque[10], A. Mobasheri[11], L. Paolozzi[12], F. Petit-Dop[12], J. Sellam[5], W. E. van Spil[2], H. Weinans[2], C. Ladel[4], J. Loughlin[1], J. Bacardit (Jaume.Bacardit@Newcastle.ac.uk)[1], for the APPROACH consortium

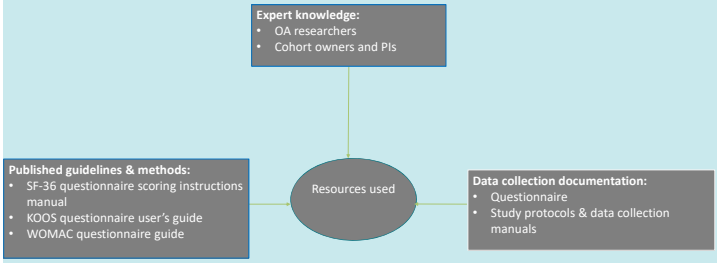
[1]Newcastle Univ., Newcastle upon Tyne, United Kingdom, [2]Univ.ir Medisch Centrum Utrecht, Utrecht, Netherlands, [3]Lygature, Utrecht, Netherlands, [4]Merck, Darmstadt, Germany, [5]Hosp. Saint-Antoine, Paris, France, [6]Servizo Galego de Saude, A Coruña, Spain, [7]Diakonhjemmet Hosp. AS, Oslo, Norway, [8]Leids Univ.ir Medisch Centrum, Leiden, Netherlands, [9]GlaxoSmithKline, Philadelphia, PA, [10]AbbVie, Chicago, IL, [11]Univ. of Surrey, Guildford, United Kingdom, [12]Inst. de Recherches Servier, Suresnes, France

Background

- APPROACH is an Innovative Medicines Initiative/European Union funded consortium of scientists and clinicians from academia and industry
- APPROACH aims to use biomarkers to assess which patients are at risk of knee osteoarthritis (OA) and most likely to progress in a 2 year timeframe and prioritise them for inclusion to OA clinical trials
- The consortium will recruit subjects to APPROACH from existing OA cohorts using machine learning algorithms, hence harmonisation of the data is needed so we can use common models on different cohorts

Inherent structural differences

- Cross-population data harmonisation is needed to align data from the different OA cohorts prior to application of algorithms to recruit patients
- This process involves transforming the data from multiple sources
- This can be challenging due to variability in data collection protocols used resulting in inherent structural differences (ISD) in datasets



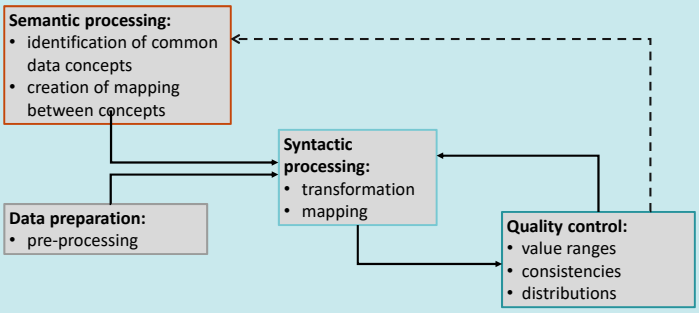
Methods

- We obtained data from five European OA cohorts:
  CHECK, DIGICOD, PROCOAC, MUST and HOSTAS
- The identified ISDs were categorised into semantics, syntactics and transformation differences.
- Developed specific strategies to address each type of difference identified based on questionnaire guidelines and expert knowledge.

- Semantics: conceptually similar data elements that could be mapped to common higher-level data elements, e.g. 'pain present on knee flexion' and 'pain present on knee extension' were mapped to 'pain on movement of knee'
- Syntactics – mapping : differences in coding practices between datasets, e.g. {1=positive, 2=negative} vs. {1=no, 2=yes}
- Syntactics - transformations: manipulations of representation schemes, where functions were employed to map from one type to another e.g. 'bodily pain score' = round (mean((6 - SF7) * 20, (5 - SF8) * 25)), where SF7 = 'bodily pain in the past 4 weeks' with values ranging from 1 = 'no pain' to 6 = 'very severe' and SF8 = 'pain interference in your work in the past 4 weeks' with values ranging from 1 = 'not at all' to 5='extremely'.

Harmonisation Pipeline

- The mapping rules were further executed as an automated software pipeline
- Quality control methods were employed to identify errors in patterns and outliers in the harmonised datasets



Results

- Our harmonisation process resulted in eight different datasets. For each cohort, we have one harmonised version of CHECK to train machine learning models and the actual cohort data to select patients
- Over 5520 records were harmonised across the five cohorts
- Overall, harmonisation of the demographic variables was easiest, and due to variations in data collection protocols, harmonisation of lifestyle and radiographic variables was most difficult

Dataset	No. of variables	No. of variables mapped
MUST	884	80
HOSTAS	124	50
DIGICOD	772	57
PROCOAC	301	53

Conclusions

- Data harmonisation was used to facilitate recruitment for APPROACH
- Transforming the data from existing cohorts to a common set of variables enabled the use of machine learning technology to identify fast OA progressors
- We believe this strategy will enhance the power and efficiency of OA clinical trials

The research leading to these results has received support from the Innovative Medicines Initiative Joint Undertaking under grant Agreement n° 115770, resources of which are composed of financial contribution from the European Union's Seventh Framework Programme (FP7/2007-2013) and EFPIA companies' in kind contribution. See IMI Europe.