

Better and more relevant scientific results can be obtained when analyses are performed on larger, more diverse datasets. This is because it is possible to detect smaller effects and check whether results still apply in different settings. Bringing together existing datasets for analysis is not always possible, because the data can be sensitive or permission has not been given to share them.

DataSHIELD (Data Aggregation Through Anonymous Summary-statistics from Harmonised Individual-level Databases) is a software system that allows users to analyse data at source by giving them the ability to run controlled commands on the data. The results returned by the commands are designed not to reveal any of the underlying data, only summary information that is still useful for analyses.

Survival analyses allow understanding of the time taken for individuals to undergo an event such as death or disease. The effects of different behaviours (e.g. smoking) and characteristics (e.g. body weight) on survival times can be studied, so that we can learn how we might improve survival times. Our work provides new functionality for DataSHIELD that allows survival analyses to be run on multiple, distributed datasets, while aiming to prevent access to the underlying data. This will enable new analyses to be run across datasets that were not previously possible.