Abstract

There is an abundance of data and most of this data is not collected with the aim of usage for health research. Single sources can provide information but often the data becomes more valuable for health research after linkage with other data.

For cardiovascular disease research there are multiple well known data sources. There are 3 different sources of data: Population based, hospital based and disease based and the phenotype depth and the sample size of the data varies between the sources. For example registries and electronic health records have less phenotype depth compared to multi-omics and imaging data sources, however their sample size is much larger.
This means that not every data source can be used for every purpose and research question and it is therefore important to identify which data sources you need to use and link in order to address your research question.

Most countries have nationwide registries, including hospital discharge information, vital statistics, tax information, population register and some countries also have nationwide disease registers such as SwedHF which is a nationwide heart failure register in Sweden. These registries can provide valuable information. For example real world incidence estimates, trend analysis and because of the large sample size it allows for sub-group analysis which is often limited in cohort studies.

There are also international disease registries for cardiovascular disease. Within these registries multiple sites from a country contribute to the registry. The value of such registries is that is provides insight in contemporary patterns in disease management, guideline implementation in daily practice and identifies potential barriers for optimal management.

The value of EHR is better insight in real-world patient, and the information in open text may have great potential. Furthermore, initiatives are currently undertaken to use EHR for the selection of trial population and this is looking promising and may save time and money.

Two datasets on itself cannot always be used to address a certain research question however when you link the two dataset the combination of the two sets can become of use. Furthermore, linkage can be an efficient method for follow-up of large groups of people, and specific sub-groups, it can be used to enrich cohorts and disease registries, it can be used to identify new risk factors, for example when new information can be obtained from open text using text mining and machine learning can be applied to these big data. And when there is feedback in the EHR of the new information obtained from linkage this can support clinical decision making.

However before linkage can become valuable we first need to overcome some challenges. An important challenge is the privacy issue. How to do the linkage in such way that an individual can be identified in the various data sources without violating the privacy legislation. The Turing institute describes the most important steps for privacy preserving datamining and this document can be recommended. An elegant method that is increasingly used is the federated method because with the use of algorithms you can analyze data hosted at different locations while the data remains at their location.

Another important challenge is that the data from the various sources are not always directly usable, for example open text first needs to be mined and they are not always directly compatible.
Furthermore there is variation in the ability to link to an individual and there is variation in the quantity of data and importantly there is variation in the definition of determinants and outcome.

Therefore data harmonization is utterly important. But this can be challenging especially when we consider phenotyping. For example Heart failure, some data sources only report HF but don't define HFpEF and HFrEF or have different definitions. Considering that the use of data is only increasing it would be good to incorporate this in the guidelines and for example provide definitions in commonly used harmonization language such as OMOP.

In this presentation some examples of data linkage projects will be shown.