EU-STANDS4PM
standards for in silico models for personalised medicine

# A European standardization framework for data integration and data-driven in silico models for personalised medicine – EU-STANDS4PM

## EU-wide mapping report with focus on international databases, collections and registries

Deliverable 1.1

# Imprint


## Authors

Alphabetical: Ali Manouchehirnia[1], Arshiya Merchant[2], Ingrid Kockum[1], Niklas Blomberg[2]

[1]Karolinska Institute, Sweden

[2]ELIXIR, EMBL-EBI, United Kingdom



## Publisher

EU-STANDS4PM administrative office on behalf of the EU-STANDS4PM consortium, June 2022 Forschungszentrum Jülich GmbH, Project Management Jülich, Germany

Contact: Marc Kirschner (m.kirschner@fz-juelich.de)

# TABLE OF CONTENTS

# Introduction

## What is personalised medicine?

Medicine has always been personal, but the extent of personalisation for each individual has varied. The evolution of DNA sequencing, along with its dramatically declining cost, has resulted in an increased understanding of genomic variants in human health and subsequently a greater degree of personalisation within healthcare (1, 2).

Personalised Medicine (PM), also termed precision medicine and sometimes genomic medicine, represents an innovative approach to the provision of healthcare that cuts across various sectors including (but not limited to) research communities, industry, funding and regulatory bodies, and nation healthcare systems (3). While there is no universal definition of PM, inspired by the Horizon 2020 Advisory Group for Health (4) we adopt the following definition: Personalised Medicine refers to a model using characterisation of individuals' phenotypes and genotypes (e.g. molecular profiles, medical imaging, medical history, and lifestyle data) to tailor the right intervention strategy for the right person at the right time, and/or determine the predisposition of disease, and/or to deliver timely and targeted prevention . PM holds real promise for improving both individual and overall population health, and the sustainability of healthcare systems across the globe (4). The full potential of PM can be realised through different approaches, including the stratification of patients into subgroups using specific markers and the integration of omics and lifestyle data to enable *in silico* modelling to support clinical decision-making (4).

## What are data sources and how are they fundamental for personalised medicine?

Central to life science research, and indeed, PM, is the collection, analysis, and implementation of 'data'. In this context we are defining 'data' to mean any data relevant for developing PM methods. In the last four decades, there has been an exponential growth in life sciences data, resulting in the creation of thousands of data resources to store, curate, and share this data (5, 6). There are a large number of established data repositories, collections, and centres that are in existence currently and are accessible online. These data sources vary in the scope of data housed, ranging from ecological studies, oceanographic expeditions, gene sequences, protein structures, toxicological assays, and more (7). Data sources in the context of this report encompass any dataset (or project) that provide data that is relevant to the personalisation of medicine. The aim of this report is to give an overview of which types of data are relevant for PM and what standards are used for different types of data relevant for PM.

# Types of data sources

## Data repositories

A data repository, also called data archive or data library, is a general term used to refer to a destination designated for data storage. The purpose of a data repository is to keep a certain population of data isolated so that it can be mined for analytical purposes at a later date. Overall, data

repositories support analytical methods and modelling approaches to help reach the goal of PM. There are a variety of data repositories, this report will focus on reference databases and cohort data.

## Reference databases

Reference databases are a type of established science data repositories. These data resources represent a type of research infrastructure which is virtual and distributed and researchers across the globe can access them without needing consent. Reference databases consist of aggregated and harmonised data from a large range of projects that make summary data available for the wider scientific community. They span across different types of biological data, including genomic, expression, sequence data, proteomics, metabolomic, epigenetic, and microbiome. Depending on the type of database, it houses carefully curated datasets or takes depositions from wider communities and can be closed or open access. For examples of reference databases please refer to Table 1.

Reference databases can be used for various purposes such as validating research or comparing a particular biomarker across different populations. They have been used in a research setting for the past two decades and play a critical role in health research as they are responsible for ensuring its quality and reproducibility (8). Recognising the importance of reference databases and their value in health research especially, almost all funders and scientific journals now strongly recommend or require deposition of research data into open access data resources. These requirements are however often in conflict with core privacy principles of EU (9). Most of the data used in PM is by nature personal, e.g. medical history and genetic data, and thus cannot according to GDPR be shared even in pseudonymized form without adequate safeguards such as giving the research subjects rights to control the processing of their data including withdrawal of data. This is in conflict with how most data repositories operate currently. A further problem with sharing of data is that according to GDPR sharing of personal data outside the European Economic area requires use of the European Commission standard contractual clauses accompanied by thorough legal assessments which is currently complicating sharing outside Europe (10). Specifically the current EU standard clauses are in conflict with US federal law making sharing of data with researchers from US complicated (10).

## Case-control data

Case-control studies are retrospective observational studies in which individuals with the outcome (cases) and without the outcome (controls) are clearly defined at the time of study. Case-control studies retrospectively assess the statistically significant difference in the rates of exposure to a defined risk factor(s) between the cases and controls. Case-control studies should include cases and controls that are identical except for their outcome which often is disease status. Case-control studies can suggest associations between the risk factor and development of the disease, however, no causality can be drawn.

Selection of cases and controls is probably the most important step in conduct of a case-control study. Cases are often selected from reliable sources such as a disease registry and included in the study based on objective inclusion and exclusion criteria. Controls should also be selected carefully and should be matched to cases on the basis of various factors (e.g. age, sex, ethnicity) to ensure an unconfounded estimation of the associations.

Due to their retrospective design cases-control studies are more prone to biases such as selection bias, recall bias, etc than cohort studies. Nevertheless, case-control studies are cheap (compared to cohort studies) and can provide fast results, in particular for the studies of rare diseases or diseases with long latency periods. An example of a case-control study is the Environment in Multiple Sclerosis (EIMS) cohort study (11). The EIMS study recruits newly diagnosed multiple sclerosis (MS) patients and non-MS population based matched controls in Sweden. Cases and controls in the study answer an extensive life-style/environmental questionnaire and donate biological samples for further analysis. The EIMS study has been a basis for many scientific investigations on the association between environmental and genetic factors and risk of MS.

## Cohort data

Cohort data plays an essential role in medical research. A cohort study is a type of longitudinal observational study that investigates incidence, cause, and/or prognosis of disease and aims at establishing links between risk factors and disease outcomes. Cohort studies are often designed with particular research questions in mind. As related to the level and quality of evidence, cohort studies are ranked below meta-analysis and randomised controlled trials but higher than case–control studies, cross sectional studies and case series/reports.

A major strength of a cohort study is the fact that multiple outcomes can be investigated. On the other hand cohort studies can be expensive and time-consuming as for reliable findings they need to include large populations and long periods of observation for sufficient number of cases to develop. Occasionally cohort studies are the only way to explore certain outcomes, for example, when it is unethical or impractical to conduct a randomised control trial or for situations when it is not known what treatment or exposure should be used in a randomised control trial.

The landscape of cohort data is vast and varied. In addition to the aforementioned example, there are several other national or international cohorts such as the Personal Genome Project (PGP)[1], the Danish Blood Donor Study (DBDS)[2] or pan-European studies such as European Prospective Investigation into Cancer and Nutrition (EPIC)[3]. Furthermore, there are directories such as the UK Medical Research Council's Cohort Directory[4] and the EU Joint Programme - Neurodegenerative Disease Research (JPND) Global Cohort Portal[5] which is a searchable catalogue of cohort studies that covers both disease-focused and general population studies.

Data from cohort studies has been and needs to continue to be deposited in global or disease specific repositories such as Dryad or Project Datasphere, respectively, to promote a more connected data ecosystem in clinical research by applying the FAIR (Findable, Accessible, Interoperable, and Reusable) data principles[6]. Deposition of cohort data into repositories is beneficial as it provides stable, long-term housing of the data, improves the security and quality of archiving through active data curation, increases the discoverability of data through the application of metadata (i.e. data about the data) schemes, and facilitate the processes of request and transfer of data from generators to users, as well

[1] https://www.personalgenomes.org/
[2] https://www.danishnationalbiobank.com/danish-biobank-register
[3] https://epic.iarc.fr/
[4] https://mrc.ukri.org/research/facilities-and-resources-for-researchers/cohort-directory/
[5] https://www.neurodegenerationresearch.eu/
[6] https://www.dtls.nl/fair-data/personal-health-train/

as tracking data utilisation[7]. One of the biggest limitations is that many existing repositories are not adequately set-up for clinical studies, and do not fully support the forms of restricted access often required for datasets containing individual patient data[8]. One way to overcome this is to enable effective, cross-border access to data, in a coordinated, secure, federated environment that enables population-scale genomic, phenotypic, biomolecular, and clinical data to be accessible across international borders.

## Data Repository of cohort data – Federated EGA

The European Genome Phenome Archive (EGA)[9] is a resource for the permanent archiving and sharing of controlled-access genetic and phenotypic human data resulting from biomedical research projects. The EGA is one of several ELIXIR Core Data Resources and is the recommended repository for sensitive human data (8).

The EGA is in the process of becoming a federated model, which will enable data to be hosted locally at e.g. research institutes, laboratories, clinics, etc. at ELIXIR Nodes. The overall goal is to provide secure, standardised, documented and interoperable services under the framework of the EGA. The fundamental principle of the EGA federated framework is that data sets remain within appropriate jurisdictional boundaries; whereas, metadata (for example, data set descriptions) are centralised and searchable through a common application programming interface (API). After data discovery, access to the data themselves can be requested from the source, for example, by applying to a data access committee, to establish agreements for data use.

As demonstrated above, data and metadata collected by disparate cohorts varies greatly, and the information is collected for different purposes. Standardisation and interoperability of these data is critical to prevent seclusion of data in silos and can be achieved through the application of FAIR—that is findable, accessible, interoperable, and reusable—data principles, thereby benefiting the cohort owner and the wider community. ELIXIR is committed to the coordination of metadata standards for such data, for example, within the Federated Human Data Community, and ensuring alignment with international project such as CINECA[10], and international collaborations e.g. International HundredK+ Cohorts Consortium (IHCC)[11] and the Global Alliance for Genomics and Health (GA4GH)[12].

## Biobank data

The collection of biological samples is not a new concept; however, it has historically been conducted on a case-by-case basis[13]. Over the last decade, significant investments from funding agencies, improvements in technologies, and developments in standards for the collection, storage, and use of samples has resulted in the creation of sizeable biobanks that have become crucial resources for biomedical research(12, 13),[14].

[7] https://www.dtls.nl/fair-data/personal-health-train/
[8] https://www.dtls.nl/fair-data/personal-health-train/
[9] https://ega-archive.org/
[10] https://www.cineca-project.eu
[11] https://ihccglobal.org/
[12] https://www.ga4gh.org/
[13] https://www.bbmri-eric.eu/wp-content/uploads/BBMRI-Biobanks-and-the-Public.pdf
[14] https://www.bbmri-eric.eu/wp-content/uploads/BBMRI-Biobanks-and-the-Public.pdf

A biobank, in its simplest form, is a repository for biological samples and associated personal health information collected in a systematic fashion (14). Biobanks can be grouped into two categories: disease-specific or population-based. Disease-specific biobanks gather information and material from individuals who have a particular disease or condition. Tumour banks are a great example, as they can guide clinical decision-making in addition to contributing to research (15). The number of individuals included in disease-specific biobanks can vary and is typically smaller than population-based biobanks where individuals are recruited from the general population, often on the basis of their residence, and are mainly used for research purposes (16).

The UK Biobank, an example of a population-based biobank, is a national initiative that recruited 500,000 people aged between 40-69 years in 2006-2010 from across the country to take part in this project. They have undergone measures, provided blood, urine, and saliva samples for future analysis, as well as detailed information about themselves and agreed to have their health followed. Over many years this biobank will become a powerful resource to help scientists develop a better understanding of differences between disease onset between individuals. Population-based biobanks, relative to disease-specific ones, are more flexible as they are set-up to support a broad range of scientific investigations such as cross-sectional and case-controlled studies over a number of years (16). However, for rare conditions disease specific biobanks will be more useful than population-based biobanks as there will not be sufficient number of individuals with a given condition.

Biobanks have promoted a paradigm shift from a “one size fits all” approach and have laid the foundation to a more personalised approach. The availability of clinically relevant and carefully curated biological material has enabled the stratification of patients based on specific characteristics resulting in the creation of more targeted therapies with reduced side effects, development of *in silico* models, as well as improved clinical trial designs and prevention strategies (13).

However, the utility of biobanks remains questioned, as their value is linked to its content, specific practices used in the management of materials and data, and consent procedures (14). Biobanking practices widely vary both nationally and internationally, and represent a huge barrier to cross-border research and collaborations. Successfully overcoming these barriers and harmonising the fragmented European biobanking community is the Biobanking and Biomolecular Resources Research Infrastructure (BBMRI-ERIC), a pan-European research infrastructure. BBMRI-ERIC facilitates access to high-quality samples and data from over 500 biobanks and sample collections across 21 EU countries (17). The BBMRI-ERIC Directory is a useful tool that shares aggregated data about biobanks across Europe that are open to collaboration and provide access to others.

## Patient registries

Simply defined, patient registries refer to a collection of standardised information about a group of individuals who share a condition or experience (18). It is an umbrella term that includes product registries, health services registries, disease/conditions registries, and a combination of all three (18). Patient registries play an integral role within healthcare by acting as a catalyst to improve health outcomes, reduce healthcare costs, and increase the value of healthcare services (19).

## Research or clinical-led registries

Traditionally, patient registries—mainly health services, product, and/or disease registries—were researcher-generated. Clinical or research institutions used private or public funds to collect observational data for a pre-defined purpose (18). Particularly popular were disease, especially rare disease, registries, as they provided a wealth of information and key insights into the course of disease (20). Acknowledging the utility of registries, countries, especially Sweden, Denmark, Australia, UK, and USA, started setting up and funding their own national registries for various purposes including monitoring of quality of care for specific conditions (21). For example, Sweden's national registry for acute coronary care played an integral role in improving Swedish hospitals' adherence to nine interventions recommended by the European Society of Cardiology, resulting in decreasing the average thirty-day mortality rate for patients who had an acute heart attack by 65% and the one-year mortality rate by 49% (21). Patient registries are now integrated into routine clinical practice with systematic data capture through electronic health records (22). In some instances such as the Swedish Multiple Sclerosis registry they have been developed into clinical support systems (23). Another initiative are European Reference Networks (ERNs) for rare and low prevalent complex diseases[15] which now includes 24 different networks and 25 countries.

Patient registries are a crucial resource within healthcare and research, as they enable healthcare providers to compare, identify, and adopt best practices to care and treat patients (19). Additionally, as personalised medicine further develops, the number of patients per treatment will decrease, initiating different types of collaborations between research centres. Patient registries are in a unique position to promote this collaboration and enable compilation of patient data that will withstand rigorous analysis (24).

## Patient powered registries

Another type of registries are patient powered registries (PPR). PPRs can take many forms and the definition of the term has been debated. It ranges from research-generated patient registries where patients contribute with data in the form of patient reported outcomes to registries run ("powered") by patient organisations or family members who then manage and control the collection of data, the research agenda for the data, and/or the translation and dissemination of the research from the data (18). An example of a research initiated registry where patients are now actively contributing data is the Swedish Multiple Sclerosis registry where patient-reported outcomes (PROMs) are captured using standardised questionnaires capturing different disease symptoms such as the Multiple Sclerosis Impact Scale (MSIS-29) and EuroQuol 5D (EQ5D) (23). One advantage of using PROMs is more frequent reporting than can be achieved when symptoms are reported by a physician during a clinical visit.

The earliest documented advocacy network run by PPR is the Hereditary Disease Foundation, initiated in 1983 with the goal of providing updated information to patients suffering from Huntington Disease and their families as well as collecting samples to advance the research[16]. In 2012 it was estimated that 45% of disease advocacy organisations supported PPRs or biobanks (25). PPRs run by patient organisations are sometimes used as a way of recruiting patients into clinical trials, an example being

[15] https://webgate.ec.europa.eu/ern/
[16] https://www.hdfoundation.org/

TMJ association[17].There have been concerns raised regarding quality of data in PPRs, it is likely that there is a bias in recruitment resulting in over representation of educated and motivated patients and families. In addition, the standardisation of data collection is usually less mature than in research or clinical-led registries but still in many cases valuable information not available anywhere else can be found in PPRs, this is especially true for rare diseases (18).

## Clinical trial registries

A clinical trial registry is a platform that catalogues clinical trials and its respective results. They contain metadata from clinical studies and occasionally (aggregate) analysis results. An example of this platform is the International Clinical Trials Registry Platform (ICTRP) launched by the World Health Organisation (WHO) in 2006. ICTRP has 17 data providers, such as ClinicalTrials.gov and EU Clinical Trials Register (EU-CTR).

There are many different types of clinical trials, with varied structures, and can be registered by different authorities such as the principle investigator, trial sponsor (eg. pharmaceutical company) or trial organiser (eg. a contract research organisation). Registration of clinical trials and submission of corresponding results is mandatory in accordance with Section 801 of the Food and Drug Administration Amendments Act (FDAAA 801) and Regulation (EU) No 536/2014). Clinical trial registries serve to increase transparency and support unbiased reporting of trial results, as well as identify gaps and prevent unnecessary duplication (26). Additionally, they also help to coordinate multinational clinical trials and aim to facilitate recruitment by making healthcare providers and potential participants aware of actively recruiting trials[18]. Deposition of clinical trial data (similar to cohort data) into repositories can serve to safely and effectively share data from these studies.

## Concerns about clinical trial registries

Overall, clinical trial registries are an extremely useful tool to help further personalised medicine. However, there remain challenges in their creations and/or use, including a lack of: coordination between national and international initiatives, harmonised data structures, data sharing and transparency, as well as sustainability[19].

Attempting to address the lack of coordination, data sharing, and transparency WHO launched the ICTRP to make information about all clinical trials involving human beings across the globe publicly available (27).

Simultaneously, the European Clinical Research Infrastructure Network (ECRIN) is addressing the lack of standardised protocols and data structures, as well as sustainability of patient registries, specifically clinical trials registries. ECRIN links scientific partners and networks across Europe in order to facilitate multinational clinical research. ECRIN's scientific partners actively work to develop shared tools, procedures, and practices to facilitate multicentre studies and manage outcomes data. ECRIN's Metadata Repository (MDR), largely based on data from trial registries and PubMed, is a tool that greatly facilitates the identification and access to data from clinical trials that have ended by

[17] http://www.tmj.org/
[18] https://www.who.int/clinical-trials-registry-platform
[19] https://www.ema.europa.eu/en/human-regulatory/post-authorisation/patient-registries#

centralising the information and providing a single point of access[20]. The MDR also distinguishes registry results entries from the original registration and gives each entry a separate link. Data is collected and aggregated and presented in a searchable form, to allow a broad array of uses.

In addition, although patient registries are able to capture data that is reflective of the "real-world" and representative of the patient population. However, interpretation of the data requires analytic methodology that addresses the bias that is present within observational studies.

### Administrative health data

In the absence of cohort data administrative health data offers a cheap alternative to cohort data. The advantages of administrative health data include large numbers, often population level coverage, minimal data loss, and systematic collection of data over long periods of time. Since administrative data is generated through the routine administration of health care and for purposes related to payment, reimbursement, caution should be taken when administrative health data is used to answer research questions. For example, in many of the Nordic countries it is possible use the social security number to link information in different administrative health registries such as the National Patient Register providing details of inpatient and outpatient diagnoses as well as medical procedures, Prescribed Drug Register providing detailed information on pharmacological treatments and Longitudinal integrated database for health insurance and labour market studies providing health insurance, parental insurance, and unemployment insurance which also if connected to molecular data provides excellent source for personalised medicine research.

### Adverse events database

Adverse reactions are a significant cause of morbidity and mortality. While these events cannot be eliminated, they can be minimised by tailoring treatments to individuals. Adverse events databases such as SIDER 4.1: Side Effect Resource[21], US FDA Adverse Event Reporting System (FAERS)[22], and the European Database of Suspected Adverse Drug Reactions[23] contain information on marketed medicines and devices and their recorded adverse drug reactions. These databases help refine current pharmacovigilance strategies to better personalise drug and device treatment by helping the general public, as well as regulatory authorities to monitor the safety of a device or active substance.

There is no comprehensive database of data sources that may be useful for personalised medicine research. It is also unclear what type of data is available in different types of studies. Hence we designed an internet based survey with the aim of characterising different data sources relevant for personalised medicine.

[20] https://ecrin.org/clinical-research-metadata-repository
[21] http://sideeffects.embl.de/
[22] https://www.fda.gov/drugs/surveillance/questions-and-answers-fdas-adverse-event-reporting-system-faers
[23] https://www.adrreports.eu/

# EU-STANDS4PM survey results

To characterise the opportunities and challenges for data driven Personalised Medicine analysis and modelling we designed a survey covering the relevant data resources. The survey consisted of 92 questions of which 52 concerned data and data standards relevant for PM and are analysed in this deliverable. These questions included context-dependent drill-downs to capture specific details for each data resource (see appendix I). The remaining questions concerning, e.g., data access and modelling are analysed in WP2-4. The survey was distributed through the network of contacts within EU-STANDS 4PM. There were a total of 71 respondents to the survey, out of which 77% (n=55) indicated that they were aware of potential data source(s) relevant to personalised medicine. Responses were predominantly covering data sources coordinated within EU member states (88%), with 33% of the respondents were involved in EU funded projects/initiatives.

## Type of data source

The most common type of datasets (or projects) (n=51) was cohort studies (39%), followed by consortium studies (18%), Biobanks (8%), general (12%) and disease specific registries (6%), Health Administrative data (2%) and other types of datasets (16%). The number of individuals in the datasets (or projects) ranged from 2 - 27 million with an average of 5 million/dataset. The largest datasets are registry or biobank based. Small datasets are dominated by cohort studies as can be seen in Figure 1.

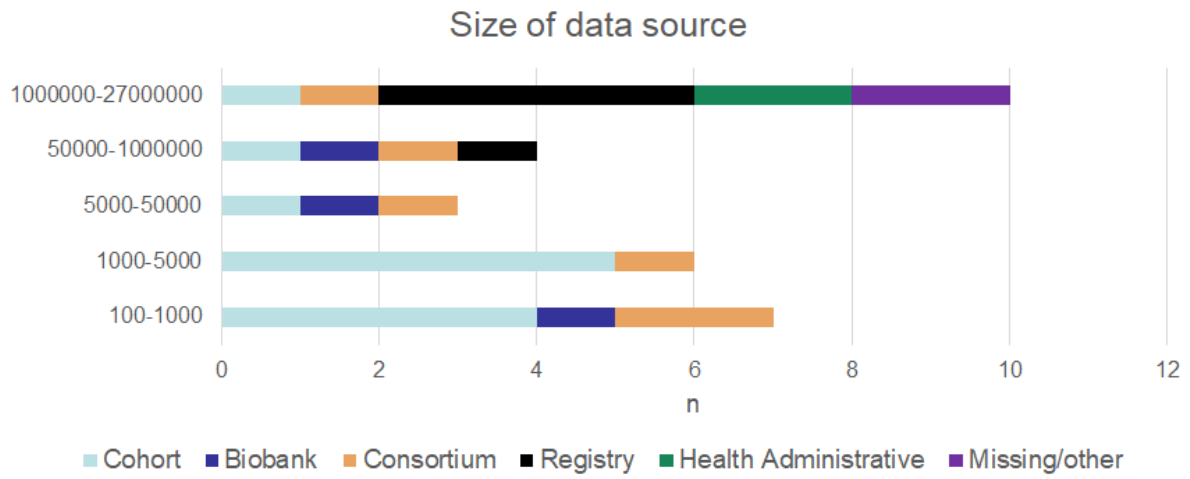



**Figure 1. Size of data source divided by type of data source.**

The primary research question that the data was gathered to answer was to identify disease risk or progression (25%), followed by prediction of disease onset (20%) and determination of disease targets and biomarkers (12%), although most datasets were collected to answer multiple questions.

Commonly data was generated within mixed research and clinical settings (44%), research studies (30%) or clinicals setting (17%). Standard formats or format guidelines of the data was highly variable and only reported for 48 of the 71 respondents. ICD-10 was the most commonly reported standard (40%), followed by “Other” (29%), “None available” (15%), and SNOMED-CT (4%). Within the “Other” free text section, respondents stated that ICD 9 was widely used. Free text responses also included mentions of OpenEHR (Electronic Health Records), an open source software available to use consisting

of open specifications, clinical models and software that can be used to create standards and build information and interoperability solutions for healthcare. DICOM, which is the international standard to transmit, store, retrieve, print, process, and display medical imaging data is also used in several instances.

## Type of data collected

Demographic information has been collected in most of the datasets (74%). Sex (83%), date of birth (69%) is the most common demographic data collected, followed by place of residence and ethnicity (both 37%).

Biomolecular data that is gathered in 77% of the respondents. The most common type of data is genotype (65%) and sequence data (56%), followed by expression (55%) and epigenetic data (31%). It is notable that this type of data is mainly available in consortium and cohort studies (Figure 2).

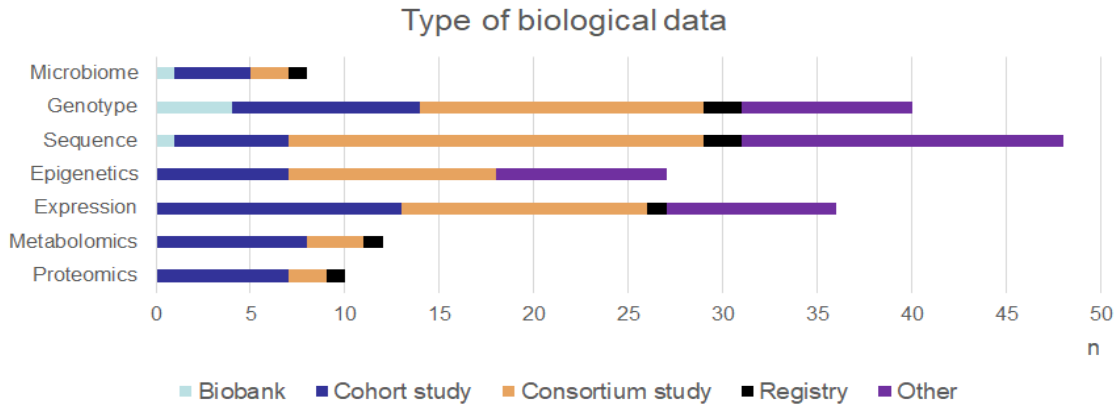



**Figure 2. Type of biological data collected divided by type of data source.**

Metadata such as data source, methodology of data generation, quality control and description of pre-analytic steps are important when reusing data and combining data from different sources. Remarkably, such data is often not available, especially for proteomic and metabolomic data, as shown in Figure 3.

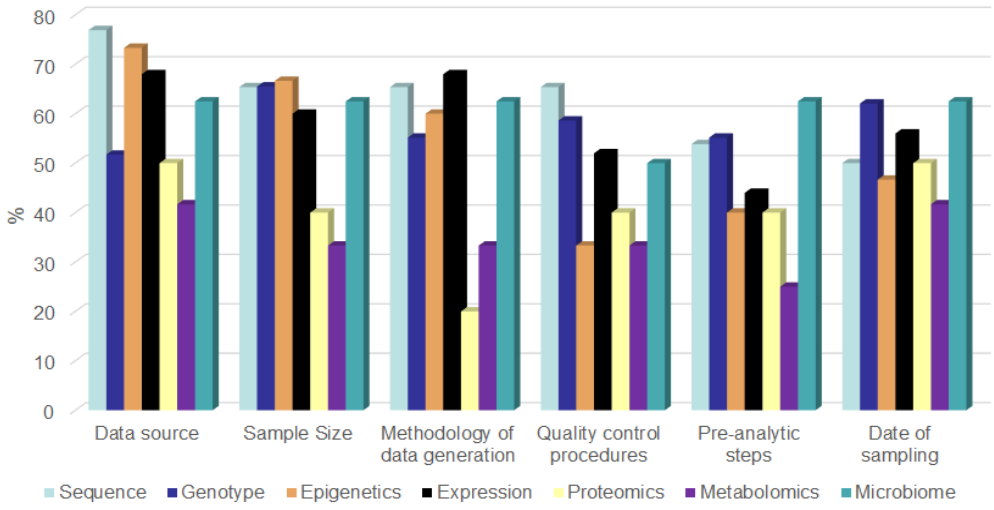



**Figure 3. Type of metadata collected divided by type of biological data.**

Traits (e.g. case/control status or healthy/non-healthy) are being collected in 60% of the studies, general clinical information and disease specific clinical information is collected in 66% and 67% of the studies respectively. Comorbidity data is mainly specialist physician (43%), registry (34%) or electronic medical record (26%) and rarely from general practitioners (11%). Distribution of use of these different sources for comorbidity is similar between different types of studies although general practitioners are only used as a source in cohort and registry data (Figure 4).

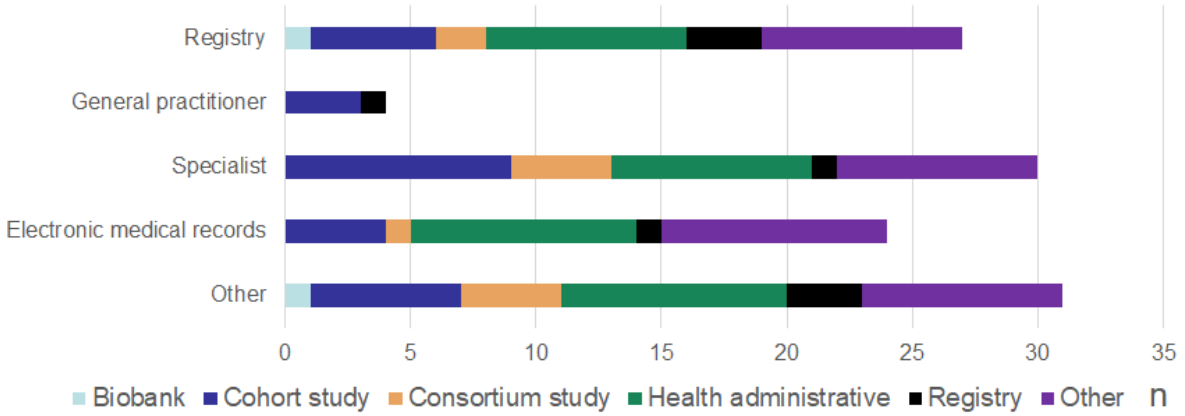



**Figure 4. Source of data on comorbidities divided by data source.**

ICD standards are used roughly to the same extent in all types of studies, but in less than 35% of them to record co-morbidity.

Data on medication is mainly collected from clinical specialists (37%) or registries (34%). The ATC standard for medication data is used in less than 32% of the studies and mainly found in in biobanks (32%) and health and administrative data sources (32%).

The main source of data on hospitalisation is from electronic medical records (28%) followed by clinical specialists (25%) and registries (25%). ICD standards are used mainly in biobank and health and administrative studies for hospitalisation data (33% for both).The most common disease specific data to collect is date of disease onset followed by disease specific treatment (Figure 5).

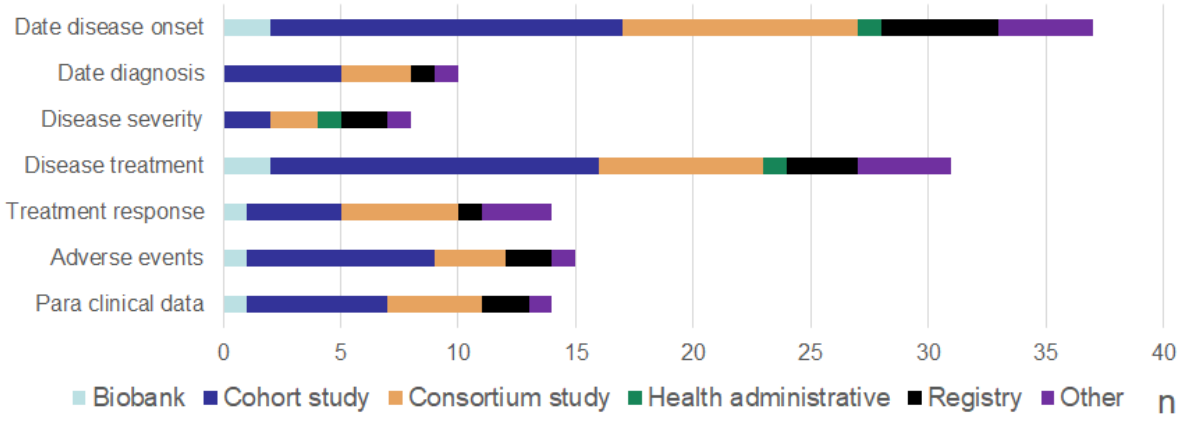



**Figure 5. Type of disease specific data collected divided by data source.**

Lifestyle and environmental exposure data was collected in only 36% data sources. The most common exposure to be captured in these studies was smoking 95%, body weight (79%) and physical activity (68%, Figure 6).

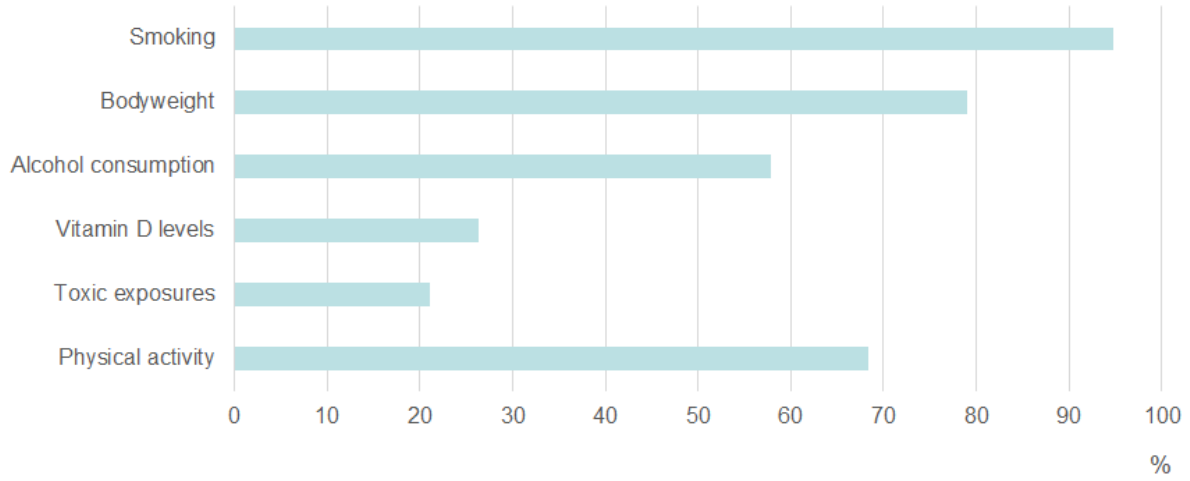



**Figure 6. Types of environmental/lifestyle exposures that are gathered.**

# Discussion

Europe is host to a wealth of data resources relevant to development of PM, for instance the GWAS catalogue[24] summarises data from over 5400 case-control studies and catalogues of biobanks, specialised cohort studies and disease registries have been constructed in many European collaboration. However, data discovery remains a major challenge: currently, a comprehensive catalogue of data sources that could be relevant to further personalised medicine research does not exist. Additionally, the type of data available within relevant sources is often unclear and to the best of our knowledge there is no study available on the limitations of the different studies for personalised medicine. The EU-STANDS4PM survey results reported here, while only constituting a limited sample, aims to fill that gap by providing a more detailed view on opportunities and limitations. Cohort data was the largest source of data in the survey, with fewer responses covering biobanks, registries, and hospital administration as sources of data.

Contrasting cohort data, only 2% of participants cited health administrative data as a sources relevant to personalised medicine. Health data such as health administrative data is often housed in healthcare facilities and cannot leave its jurisdictional boundaries, making it increasingly difficult to discover, access, and use for research purposes. However, such data has the advantage that it is usually covering a large proportion of the population and usually represents the patient population well. These datasets play a vital role in health research, as they can be used to conduct research for multiple therapeutic areas and are able to collect large volumes of data on health, social, economic, demographic, and other data. This information can be used to identify patterns, understand the impact of diet, environment, income, access to healthcare to truly personalise healthcare. While the sample in this survey is limited we surmise that health administrative data is underutilised for personalised medicine studies and future actions to mobilise such datasets for development of diagnosis, treatments and understanding outcomes between countries and regions would be important for development of PM in Europe.

There is a distinct lack of defined health data standards that are widely used in our survey. This is consistent with many other reports on health and personalised medicine data (28). The development of standards is done on a need basis rather than proactive planning for the management of health information. There is a competition between standard development organisations (SDOs) which results from the natural evolution and expanding scope of the work. This competition forces implementers to choose among multiple options and requires an additional step of mapping between standards using an interface engine for interoperability when combining data from different sources. With regards to standard formats, ICD codes, in particular ICD10 and ICD9 were listed as being most widely used. A noteworthy other format used was openEHR, an open source software that is freely available and easy to adapt. It consists of open specifications, clinical models, and software that can be used to create standards, and build information and interoperability solutions for healthcare. Standards are also well developed and widely used in the field of genomics data. Medical imaging data is quite hard to capture, and the current standard being used is DICOM, which is the international standard to transmit, store, retrieve, print, process, and display medical imaging data is also used in several instances. [ note that standards are covering different sources and modalities, useful overview

[24] https://www.ebi.ac.uk/gwas/home

in IMI EDHEN deliverable; as more and more data is collected in clinical settings standards such as OMOP (epi) and HL7 FHIR (EHR exchange) will become important, note the challenge of training and implement convergent practices in a very large number of research performing organisations across Europe]

Most of the data resources in our survey contain data on biomolecular profiling such as genotyping, soluble biomarkers, proteomics or metabolomics data. However, it is well established that biomolecular profiles can change markedly during sample collection, storage and pre-processing, limiting reproducibility and comparability of studies and collection (29). Variability may also make outcomes and diagnostics unreliable as the results from studies are not translatable to a routine clinical setting (30). Hence it is notable that many, if not most, of the studies in our survey report that data on methods of data generation, quality control procedures and pre-analytical steps is missing. European projects such as SPIDIA4P[25] has developed procedures and standards to accurately capture such aspects, the widespread adoption of these standards and recommendations across European research performing organisations would be important for reproducible research and for generating large, comparable and high quality datasets for e.g. artificial intelligence based biomarker development.

In the data sources covered in our survey there is little information tracking the treatment regimens patients' were prescribed, and tracking of their response to these treatments (clinically, physically and on a molecular level). One of the main goals in PM is to predict which individual should be given which treatment. Such predictions would come from modelling responses to treatment. Lack of data on response to treatment is hence a huge barrier to fully achieve PM. For many of the existing data sources it was not even captured which treatment patients had been given, for how long or which dose of treatment was used. This type of data is obviously central for PM. Often the available data seems more suitable for modelling different biological processes which can be indirectly relevant for PM but do not directly address response to different treatments and therefore is quite far from being useful in clinical practice. The lack of this type of data in the surveyed studies could be addressed by using administrative data and electronic health records. Similarly, there was a mixed response for health status information, and this would be a point to note going forward. Lifestyle and environmental exposures are only captured in a minority of the surveyed studies. Exposures such as smoking, body mass index are likely to greatly impact the outcome of many diseases and are therefore relevant in PM.

[25] https://www.spidia.eu/

# Recommendations

## FAIR starts with findability

Arguably the most critical part to develop the European PM data landscape is Findability - without visibility of relevant research and clinical data, however well annotated, it is lost and will not inform diagnosis and treatment. Our survey sampled but a small part of the overall relevant data resources; there is no comprehensive catalogue – or even methodology for generating such catalogues without extensive manual curation implemented on a European scale. The overall landscape is fragmented but good examples exist – for instance in EJP RD[26] the *findability support* - discoverability of catalogues, of its hosted products/metadata, provenance mechanism to trace back sources, with appropriate web search and/or machine accessible and well documented API - has proceeded through e.g. high level support for ERN catalogues, development of "Beacons" to aid identification of relevant variants for rare diseases and through the promotion of standards such as "Phenopackets" (ISO 4454).

## Foster debate through common understanding of challenges and solutions at multiple levels

The data landscape for personalised medicine is complex, sitting at the nexus between research, clinical care and health care administration. There is a lack of generic 'terminology' for the collection, analysis, and interpretation of data. There is a gap in the researchers' understanding of standards terminology and that of the policy makers (31). In order for collaboration between health data researchers, frontline caregivers, and policymakers to implement PM in healthcare systems, a common understanding of the challenges – stretching from reproducible capturing sample and data processing, through to use of common terminologies, data standards and managing trust, data protection and privacy. Standard terminology should be used in European Commission documentation to ensure a similar level of understanding and awareness for both policymakers and researchers.

## Data-driven PM requires high-quality European data sets and cohorts

Many of the data resources surveyed in this report did not capture metadata on pre-processing of samples and on data analysis. Tools, guidelines and standards to remedy this situation exist – for instance through the Spidia/Spidia4p projects. These should be implemented broadly – via training programmes and funder mandates. In addition, recommendations for capturing treatment using standard measures are available, these should be coupled to measures of disease outcome in order to better define response to treatment. However, it is important to note that while standardisation of disease codes will help interoperability, it is also the diversity of standards that supports a wide range of use-cases and contexts for data reuse (32, 33). There is a role for funding organisations to mandate this in their grants and for research performing organisations to report on how their data is used to foster new insights and improved clinical care.

[26] https://www.ejprarediseases.org/

## Accessibility to health data from different jurisdictions will promote acceleration of PM research

However, it is important to note the different jurisdictions have different needs and the health data is governed according to those needs. A federated model (similar to the FEGA) would help overcome some of the difficulties in accessing health data. The EU Health Data will help support healthcare delivery (so-called primary use of data) but also for health research and health policy making purposes (so-called secondary use of data).

In summary, Europe hosts a large number of data resources that can provide valuable insights and constitute a foundation for data driven personalised medicine. The present survey has identified a number of important limitations that prevent effective use of this large, but heterogeneous and fragmented data landscape. However, there are also many examples of good practice in the field – providing large, well annotated resources that inform data-driven personalised medicine. The challenge is to scale these individual examples of good practice into a European-level effort and build momentum towards the large multi-national data sets required for advanced analytics and learning approaches. Only by large, well annotated and diverse datasets can we overcome the risks of bias and limited applicability inherent in artificial intelligence.

# Methods

In October 2019, an invitation to complete the EU-STANDS 4PM survey was sent to the members of the consortium. The message contained a link to an electronic survey which was accessible until March 2020. The link was open and members were asked to forward the link to whoever could contribute. The survey was not anonymous, and consent was implied with completion of the survey.

The survey was developed in collaboration between members of the EU-STANDS 4PM consortium using the Survey Report program version 4.3.10.5 housed at Karolinska Institutet. The survey contained a total of 95 questions arranged in two pages. The first 59 questions were focused on the general information regarding the data source that the participant expressed familiarity with (e.g. ethical considerations, type of dataset). The remaining questions were focused on the modelling approaches being employed for the analysis of data collected in the data source. The survey was designed in an adaptive manner in which questions were only visible to the participant depending on the answers to the previous questions. For example, page two of the survey became available only if participants expressed familiarity with modelling approaches. Survey responses were analysed using descriptive statistics.

# References


1. Green ED, Guyer MS, & National Human Genome Research I (2011) Charting a course for genomic medicine from base pairs to bedside. *Nature* 470:204-213. https://www.ncbi.nlm.nih.gov/pubmed/21307933

2. Manolio TA*, et al.* (2019) Opportunities, resources, and techniques for implementing genomics in clinical care. *Lancet* 394:511-520. https://www.ncbi.nlm.nih.gov/pubmed/31395439

3. PerMed (2015) Shaping Europe's Vision for Personalised Medicine Strategic Research and Innovation Agenda (SRIA). https://www.icpermed.eu/media/content/PerMed_SRIA.pdf

4. Horizon 2020 Advisory group for Health (2016) of the Horizon 2020 Advisory Group for Societal Challenge 1, "Health, Demographic Change and Well-being". (Brussels). https://ec.europa.eu/transparency/expert-groups-register/core/api/front/expertGroupAddtitionalInfo/23709/download

5. Gabella C, Durinx C, & Appel R (2017) Funding knowledge bases: Towards a sustainable funding model for the UniProt use case. *F1000Res* 6. https://www.ncbi.nlm.nih.gov/pubmed/29333230

6. Imker HJ (2018) 25 Years of Molecular Biology Databases: A Study of Proliferation, Impact, and Maintenance. *Frontiers in research metics and analytics* May:00018. https://doi.org/10.3389/frma.2018.00018

7. Yamamoto Y, Yamaguchi A, & Splendiani A (2018) YummyData: providing high-quality open life science data. *Database (Oxford)* 2018. https://www.ncbi.nlm.nih.gov/pubmed/29688370

8. Durinx C*, et al.* (2017) Identifying ELIXIR Core Data Resources [version 2; peer review: 2 approved]. *F1000Res* 5 (ELIXIR):2422

9. Mascalzoni D*, et al.* (2019) Are Requirements to Deposit Data in Research Repositories Compatible With the European Union's General Data Protection Regulation? *Ann Intern Med* 170:332-334. https://www.ncbi.nlm.nih.gov/pubmed/30776795

10. Bentzen HB*, et al.* (2021) Remove obstacles to sharing health data with researchers outside of the European Union. *Nat Med* 27:1329-1333. https://www.ncbi.nlm.nih.gov/pubmed/34345050

11. Hedstrom AK, Baarnhielm M, Olsson T, & Alfredsson L (2009) Tobacco smoking, but not Swedish snuff use, increases the risk of multiple sclerosis. *Neurology* 73:696-701. https://www.ncbi.nlm.nih.gov/pubmed/19720976

12. Olson JE*, et al.* (2014) Biobanks and personalized medicine. *Clin Genet* 86:50-55. https://www.ncbi.nlm.nih.gov/pubmed/24588254

13. Zika E*, et al.* (2010) Biobanks in Europe: Prospects for Harmonisation and Networking. in *JRC Scientific and Technical Reports* (European Union, Luxembourg), p EUR 24361 EN. http://www.eurosfaire.prd.fr/7pc/doc/1280153287_biobanks_eu_jrc57831.pdf

14. Policiuc L*, et al.* (2018) The foundation of personalized medicine is the establishment of biobanks and their standardization. *J BUON* 23:550-560. https://www.ncbi.nlm.nih.gov/pubmed/30003718

15. Asslaber M & Zatloukal K (2007) Biobanks: transnational, European and global networks. *Brief Funct Genomic Proteomic* 6:193-201. https://www.ncbi.nlm.nih.gov/pubmed/17916592

16. Davey Smith G*, et al.* (2005) Genetic epidemiology and public health: hope, hype, and future prospects. *Lancet* 366:1484-1498. https://www.ncbi.nlm.nih.gov/pubmed/16243094

17. Saunders G*, et al.* (2019) Leveraging European infrastructures to access 1 million human genomes by 2022. *Nat Rev Genet* 20:693-701. https://www.ncbi.nlm.nih.gov/pubmed/31455890

18. Workman T (2013) Engaging Patients in Information Sharing and Data Collection: The Role of Patient-Powered Registries and Research Networks. (Agency for Healthcare Research and Quality, Rockville (MD)). https://www.ncbi.nlm.nih.gov/books/NBK164514/

19. Anonymous (2011) National disease registries for advancing health care. *Lancet* 378:2050. https://www.ncbi.nlm.nih.gov/pubmed/22177500

20. Viviani L, Zolin A, Mehta A, & Olesen HV (2014) The European Cystic Fibrosis Society Patient Registry: valuable lessons learned on how to sustain a disease registry. *Orphanet J Rare Dis* 9:81. https://www.ncbi.nlm.nih.gov/pubmed/24908055

21. Leavy MB (2014) Registries for Evaluating Patient Outcomes. eds Gliklich RE & Dreyer NA (Agency for Healthcare Research and Quality), p EHC111

22. Cohet C (2021) ENCePP Guide on Methodological Standards in Pharmacoepidemiology. (The European Network of Centres for Pharmacoepidemiology and Pharmacovigilance). http://www.encepp.eu/standards_and_guidances/methodologicalGuide.shtml#IndividualChapters

23. Hillert J & Stawiarz L (2015) The Swedish MS registry - clinical support tool and scientific resource. *Acta Neurol Scand* 132:11-19. https://www.ncbi.nlm.nih.gov/pubmed/26046553

24. Anonymous (2013) Innovation and Patient Access to Personalised Medicine: Report from Irish Presidency Conference March 20th/21st 2013. (The European Alliance for Personalised Medicine, Brussels). https://euapm.eu/pdf/EAPM_REPORT_on_Innovation_and_Patient_Access_to_Personalised_Medicine.pdf

25. Landy DC*, et al.* (2012) How disease advocacy organizations participate in clinical research: a survey of genetic organizations. *Genet Med* 14:223-228. https://www.ncbi.nlm.nih.gov/pubmed/22261756

26. Sim I, Chan AW, Gulmezoglu AM, Evans T, & Pang T (2006) Clinical trial registration: transparency is the watchword. *Lancet* 367:1631-1633. https://www.ncbi.nlm.nih.gov/pubmed/16714166

27. Anonymous (2018) International Standards for Clinical Trial Registries. (World Health Organization, Geneva), pp CC BY-NC-SA 3.0 IGO. https://apps.who.int/iris/bitstream/handle/10665/274994/9789241514743-eng.pdf

28. Lovestone, S. (2020). The European Medical Information Framework: A novel ecosystem for sharing healthcare data across Europe. *Learning Health Systems*, *4*(2). https://doi.org/10.1002/lrh2.10214

29. Huang J*, et al.* (2021) Assessing the Preanalytical Variability of Plasma and Cerebrospinal Fluid Processing and Its Effects on Inflammation-Related Protein Biomarkers. *Mol Cell Proteomics* 20:100157. https://www.ncbi.nlm.nih.gov/pubmed/34597789

30. Muller H*, et al.* (2020) Biobanks for life sciences and personalized medicine: importance of standardization, biosafety, biosecurity, and data management. *Curr Opin Biotechnol* 65:45-51. https://www.ncbi.nlm.nih.gov/pubmed/31896493

31. Richesson RL & Krischer J (2007) Data standards in clinical research: gaps, overlaps, challenges and future directions. *J Am Med Inform Assoc* 14:687-696. https://www.ncbi.nlm.nih.gov/pubmed/17712081

32. Kees van Bochove, Emma Vos, Maxim Moinat, Sebastiaan van Sandijk, Tess Korthout, & Peyman Mohtashani. (2020). EHDEN - D4.5 - Roadmap for interoperability solutions (Version 1). Zenodo. https://doi.org/10.5281/zenodo.4474373

33. Canham, Steve, Ohmann, Christian, Boiten, Jan-Willem, Panagiotopoulou, Maria, Hughes, Nigel, David, Romain, Sanchez Pla, Alex, Maxwell, Lauren, Aerts, Jozef, Facile, Rhonda, Griffon, Nicolas, Saunders, Gary, van Bochove, Kees, & Ewbank, Jonathan. (2021). EOSC-Life Report on data standards for observational and interventional studies, and interoperability between healthcare and research data. Zenodo. https://doi.org/10.5281/zenodo.5810612

33. Xu B*, et al.* (2020) Epidemiological data from the COVID-19 outbreak, real-time case information. *Sci Data* 7:106. https://www.ncbi.nlm.nih.gov/pubmed/32210236

34. Abramowitz S*, et al.* (2018) Data Sharing in Public Health Emergencies: Anthropological and Historical Perspectives on Data Sharing during the 2014-2016 Ebola Epidemic and the 2016 Yellow Fever Epidemic. https://www.glopid-r.org/wp-content/uploads/2019/07/data-sharing-in-public-health-emergencies-yellow-fever-and-ebola.pdf

# Appendix

## Survey Questions

1. Number of respondents (in standard report)

2. Number of respondents aware of potential data source(s) suitable for personalised medicine research? (in standard report)

3. What is the type of dataset (in standard report) but should be reported as % of those who have responded to the question and those that have answered “other” should be classed into the existing groups (most of these have written combinations of listed answers). Group answers for later division when illustrating answers to other questions. Suggestion for groups:

4. Histogram of size of data source (number of individuals). Would be good to also illustrate answers to point 3 in this histogram i.e. what type of dataset.

5. What are the existing standard formats, format guidelines, ontologies etc? Here we probably need to classify those that answered other into groups (Some of them are same as the listed ones but have more than one standard) (potentially also include groups from point 3 in this plot)

6. What is the main research question/aim of the study with the collected/generated data? (part of standard output, but may need to go though “other” response to see if it can be classified into any of given answers. May also want to group those answers that are obviously relevant to precision medicine into one group to be used when dividing up results in some of the later questions)

7. Is demographic data being collected? (Part of standard output, but also divide up based on answer to point 6 grouped and in other plot grouped based on answers to point 3)

8. What demographic data is being collected? (Part of standard output, but also divide up based on answer to point 6 grouped and in other plot grouped based on answers to point 3)

9. Is biological data being collected? (Part of standard output, but also divide up based on answer to point 6 grouped and in other plot grouped based on answers to point 3)

10. Is biological data being collected? What type? (combines answers from several questions, probably do not need yes in plot/table but rather express as % of those that answer yes (Potentially divided up by answer to point 3 and 6)

11. What expression data is being collected? (part of standard output)

12. What sequence data is being collected? (part of standard output)

13. What proteomics data is being collected? (part of standard output)

14. What metabolomics data is being collected? (part of standard output)

15. What epigenetic data is being collected? (part of standard output)

16. What microbiome data is being collected? (part of standard output)

17. What type of data is being collected? (answers from several questions)

18. Other types of data being collected (based on answers to several different questions, probably should be % rather than n)

19. How is the comorbidity data being collected and from which source(Part of standard output, but also divide up based on answer to point 6 grouped and in other plot grouped based on answers to point 3 )

20. How is the medication data being collected and from which source? (Part of standard output, but also divide up based on answer to point 6 grouped and in other plot grouped based on answers to point 3)

21. How is the hospitalisation data being collected and from which source? (Part of standard output, but also divide up based on answer to point 6 grouped and in other plot grouped based on answers to point 3)

22. If data on hospitalisation is being recorded, is the reason for hospitalisation mentioned? (part of standard output)

23. If data on hospitalisation is being recorded, is the length/date of stay mentioned? (part of standard output)

24. How is data being collected? (based on answers to several questions)

25. Is disease specific clinical data being collected? (Part of standard output, but also divide up based on answer to point 6 grouped and in other plot grouped based on answers to point 3)

26. Is the date of disease onset (first disease manifestation or symptoms) being reported?

27. Is the date of diagnosis being reported? (Part of standard output, but also divide up based on answer to point 6 grouped and in other plot grouped based on answers to point 3)

28. Are the specific disease's severity measures being reported? (Part of standard output, but also divide up based on answer to point 6 grouped and in other plot grouped based on answers to point 3)

29. Are the specific disease's treatments being reported? (Part of standard output, but also divide up based on answer to point 6 grouped and in other plot grouped based on answers to point 3)

30. Is treatment response (in any form) being reported?

31. Are adverse events being recorded? (Part of standard output, but also divide up based on answer to point 6 grouped and in other plot grouped based on answers to point 3)

32. Is paraclinical data being collected? (Part of standard output, but also divide up based on answer to point 6 grouped and in other plot grouped based on answers to point 3)

33. What disease specific data is being collected? (based on several different questions, should probably be % rather than n)

34. Is data on environmental exposure and life-style being collected(Part of standard output, but also divide up based on answer to point 6 grouped and in other plot grouped based on answers to point 3)

35. What environmental exposure is collected (based on several different questions, should probably be % rather than n)

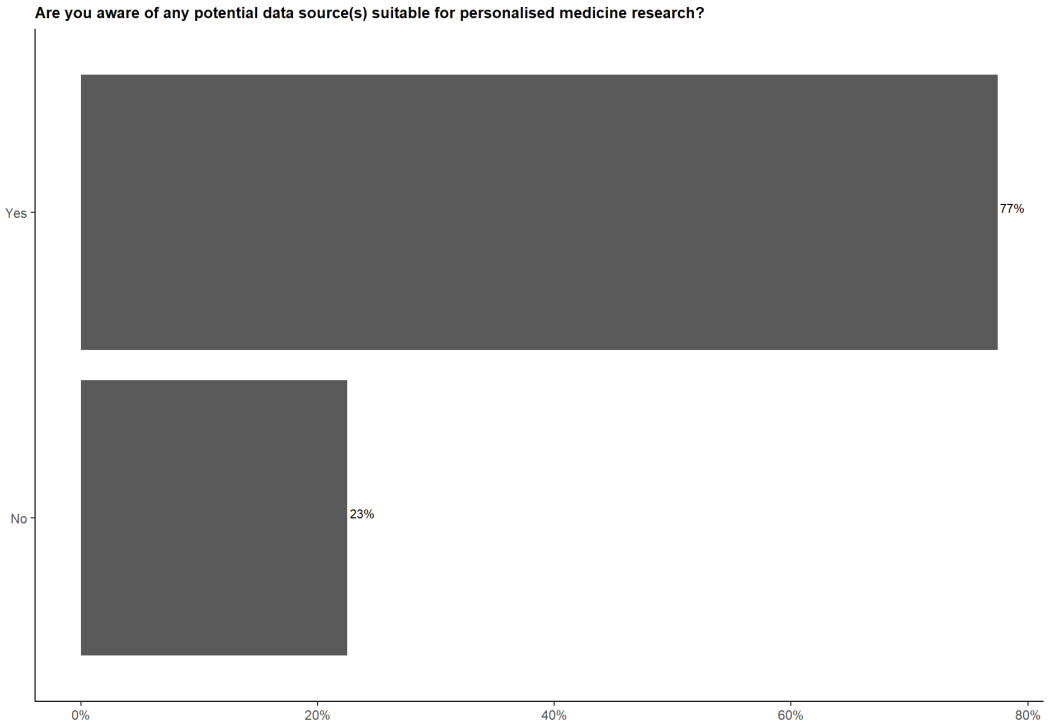
## Survey Results

**1. Number of respondents**

N=71

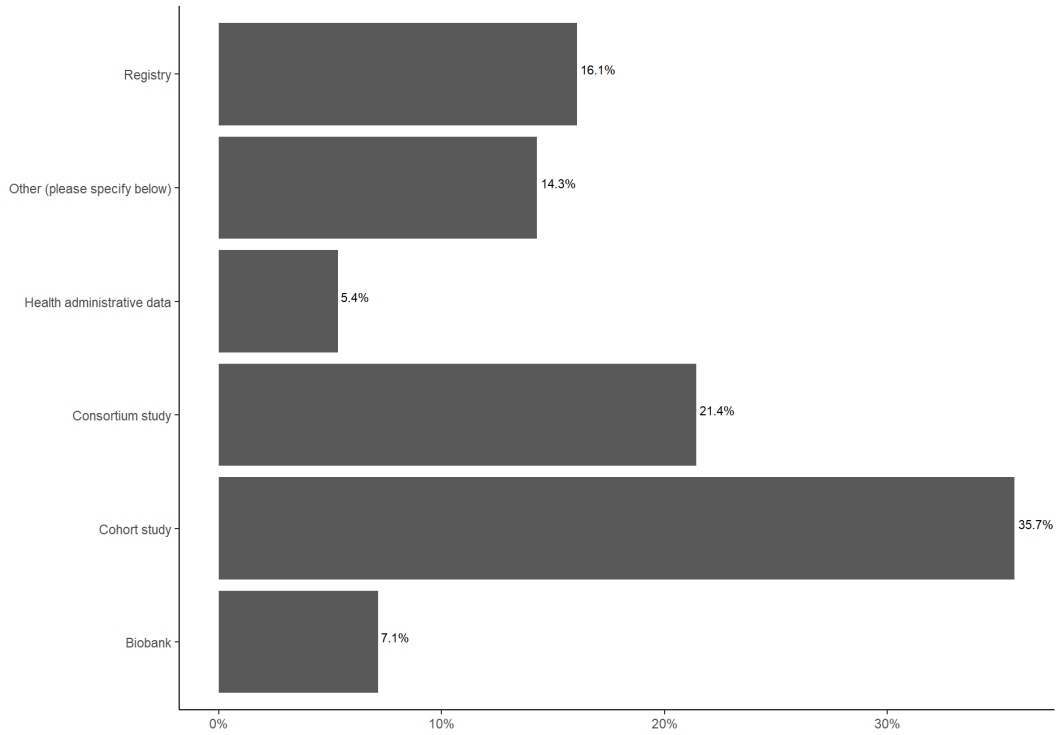
**2. Number of respondents aware of potential data source(s) suitable for personalised medicine research?**

| | Overall (n=71) |
|---|---|
| No | 16 (22.5%) |
| Yes | 55 (77.5%) |



### 3. What is the type of dataset/study?

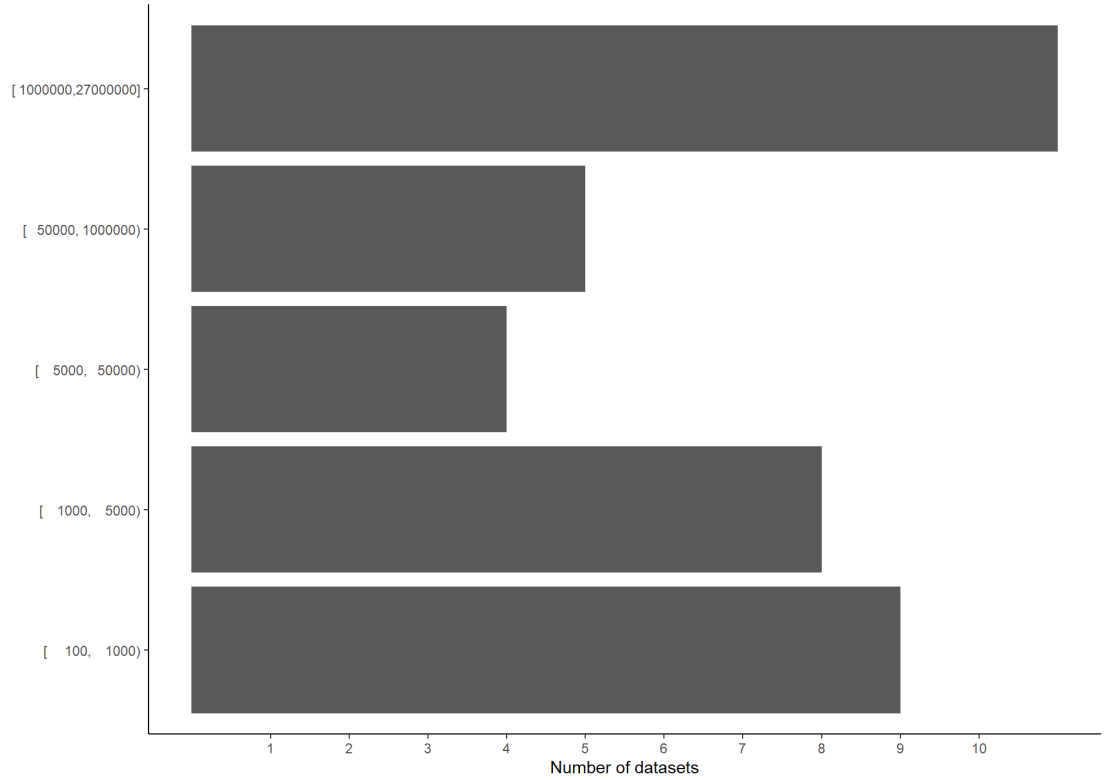
| | Overall (n=56) |
|---|---|
| Biobank | 4 (7.1%) |
| Cohort study | 20 (35.7%) |
| Consortium study | 12 (21.4%) |
| Health administrative data | 3 (5.4%) |
| Other (please specify below) | 8 (14.3%) |
| Registry | 9 (16.1%) |




Comments to the question:

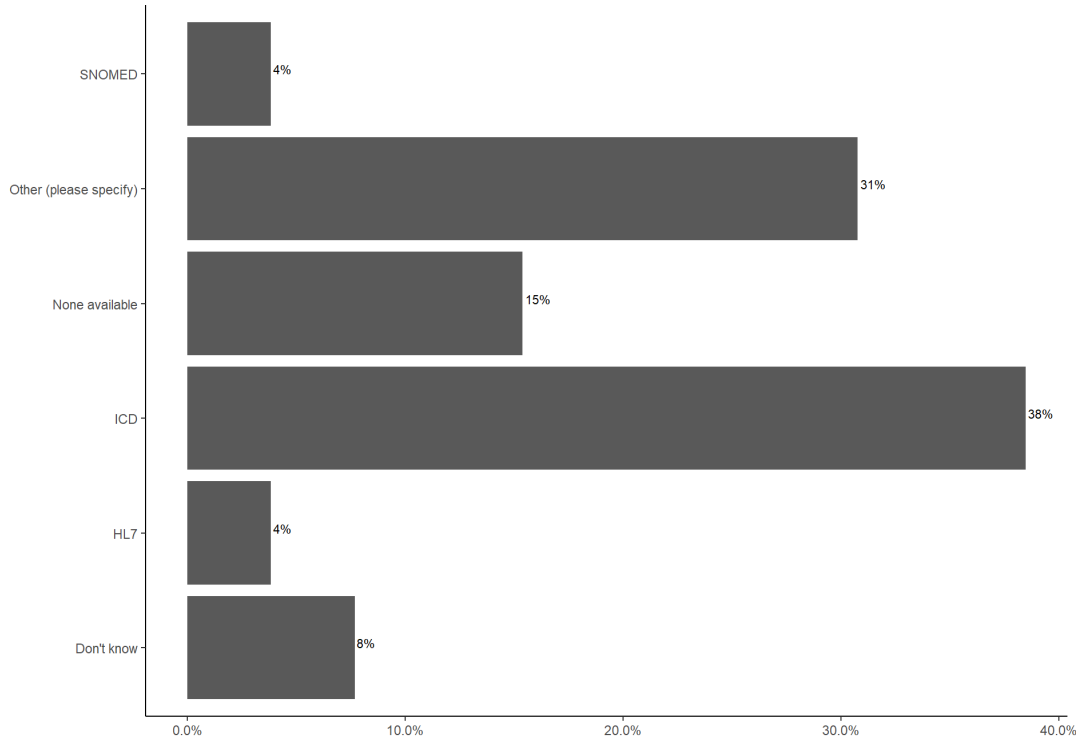
| | Overall (n=2) |
|---|---|
| **value** | |
| Several of the above answers. Only one can be entered - therefore other | 1 (50.0%) |
| Will over time cover all personal data regulated by GDPR. Behavioural, Educational, Financial, Health etc | 1 (50.0%) |

**4. Histogram of data source size.**



**5. What are the existing standard formats, format guidelines, ontologies etc?**

| | Overall (n=52) |
|---|---|
| Don't know | 4 (7.7%) |
| HL7 | 2 (3.8%) |
| ICD | 20 (38.5%) |
| None available | 8 (15.4%) |
| Other (please specify) | 16 (30.8%) |
| SNOMED | 2 (3.8%) |




For those who answered No:

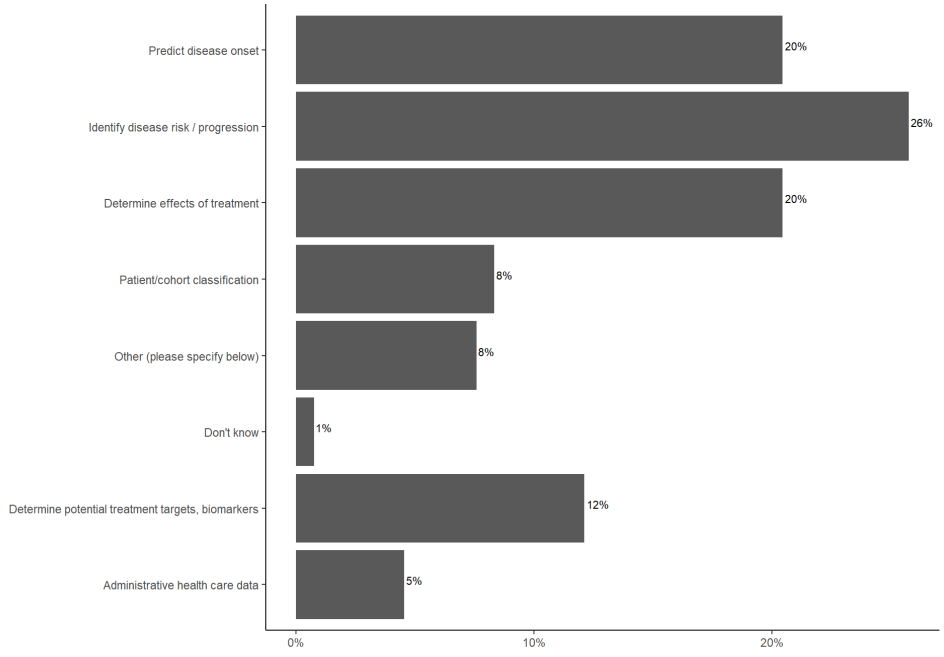
| | Overall (n=11) |
|---|---|
| **value** | |
| DICOM | 1 (9.1%) |
| I don't know | 1 (9.1%) |
| MRI | 1 (9.1%) |
| openEHR | 1 (9.1%) |
| Plink | 1 (9.1%) |
| RedCap | 1 (9.1%) |
| Several | 1 (9.1%) |
| VCF | 4 (36.4%) |

Answers divided by type of study:

| | Don't know (n=4) | HL7 (n=2) | ICD (n=22) | None available (n=10) | Other (please specify) (n=20) | SNOMED (n=2) | Overall (n=83) |
|---|---|---|---|---|---|---|---|
| **cohort_type** | | | | | | | |
| Biobank | 1 (25.0%) | 0 (0%) | 2 (9.1%) | 0 (0%) | 0 (0%) | 0 (0%) | 4 (4.8%) |
| Cohort study | 2 (50.0%) | 2 (100%) | 5 (22.7%) | 4 (40.0%) | 6 (30.0%) | 1 (50.0%) | 20 (24.1%) |
| Consortium study | 1 (25.0%) | 0 (0%) | 3 (13.6%) | 4 (40.0%) | 6 (30.0%) | 0 (0%) | 15 (18.1%) |
| Health administrative data | 0 (0%) | 0 (0%) | 3 (13.6%) | 0 (0%) | 0 (0%) | 0 (0%) | 4 (4.8%) |
| Other (please specify below) | 0 (0%) | 0 (0%) | 2 (9.1%) | 2 (20.0%) | 6 (30.0%) | 0 (0%) | 12 (14.5%) |
| Registry | 0 (0%) | 0 (0%) | 6 (27.3%) | 0 (0%) | 0 (0%) | 1 (50.0%) | 9 (10.8%) |
| Missing | 0 (0%) | 0 (0%) | 1 (4.5%) | 0 (0%) | 2 (10.0%) | 0 (0%) | 19 (22.9%) |

## 6. What is the main research question/aim of the study with the collected/generated data?

| | Overall (n=132) |
|---|---|
| Administrative health care data | 6 (4.5%) |
| Determine potential treatment targets, biomarkers | 16 (12.1%) |
| Don't know | 1 (0.8%) |
| Other (please specify below) | 10 (7.6%) |
| Patient/cohort classification | 11 (8.3%) |
| Determine effects of treatment | 27 (20.5%) |
| Identify disease risk / progression | 34 (25.8%) |
| Predict disease onset | 27 (20.5%) |



Specification of those who answered Other:

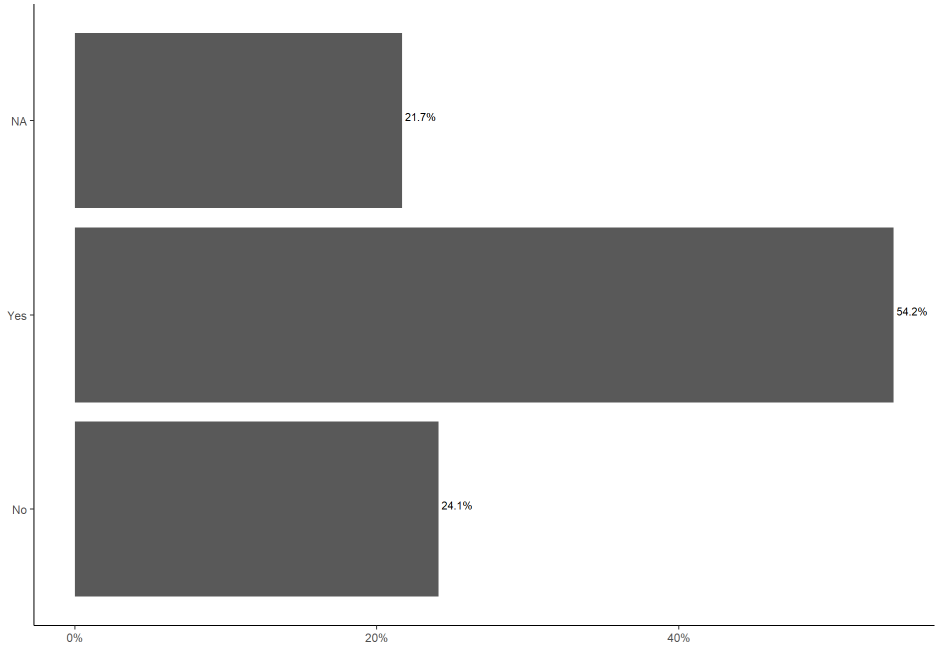
| | Overall (n=22) |
|---|---|
| **value** | |
| ALSPAC is a managed access resource which is open to investigators considering any health and social questions aiming to improve the public good. | 1 (4.5%) |
| Create a reference genome, eg. for imputation | 16 (72.7%) |
| EGA stores multiple datasets with a wide range of use-cases and consented research purposes. | 1 (4.5%) |
| Epidemiological research | 1 (4.5%) |
| More options could apply | 1 (4.5%) |
| Several of above. The purpose of the register is: statistics for the clinical units, quality improvement and research | 1 (4.5%) |
| To provide high-quality multi-omics data under open access for research | 1 (4.5%) |

Answers divided by purpose of study:

| | Biobank (n=4) | Cohort study (n=20) | Consortium study (n=15) | Health administrative data (n=4) | Other (please specify below) (n=12) | Registry (n=9) | Overall (n=83) |
|---|---|---|---|---|---|---|---|
| **data_aim** | | | | | | | |
| Determine potential treatment targets, biomarkers | 1 (25.0%) | 1 (5.0%) | 1 (6.7%) | 0 (0%) | 0 (0%) | 0 (0%) | 4 (4.8%) |
| Don't know | 1 (25.0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 1 (1.2%) |
| Multiple aims | 1 (25.0%) | 10 (50.0%) | 9 (60.0%) | 3 (75.0%) | 10 (83.3%) | 2 (22.2%) | 37 (44.6%) |
| Other (please specify below) | 0 (0%) | 2 (10.0%) | 0 (0%) | 0 (0%) | 0 (0%) | 1 (11.1%) | 3 (3.6%) |
| Determine effects of treatment | 0 (0%) | 1 (5.0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 1 (1.2%) |
| Identify disease risk / progression | 0 (0%) | 2 (10.0%) | 3 (20.0%) | 0 (0%) | 0 (0%) | 1 (11.1%) | 6 (7.2%) |
| Predict disease onset | 0 (0%) | 2 (10.0%) | 2 (13.3%) | 0 (0%) | 0 (0%) | 0 (0%) | 4 (4.8%) |
| Administrative health care data | 0 (0%) | 0 (0%) | 0 (0%) | 1 (25.0%) | 0 (0%) | 3 (33.3%) | 4 (4.8%) |
| Patient/cohort classification | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 1 (11.1%) | 1 (1.2%) |
| Missing | 1 (25.0%) | 2 (10.0%) | 0 (0%) | 0 (0%) | 2 (16.7%) | 1 (11.1%) | 22 (26.5%) |

**7. Is demographic data being collected?**

| | Overall (n=83) |
|---|---|
| **Is demographic data being collected** | |
| No | 20 (24.1%) |
| Yes | 45 (54.2%) |
| Missing | 18 (21.7%) |



Answers divided by type of study:

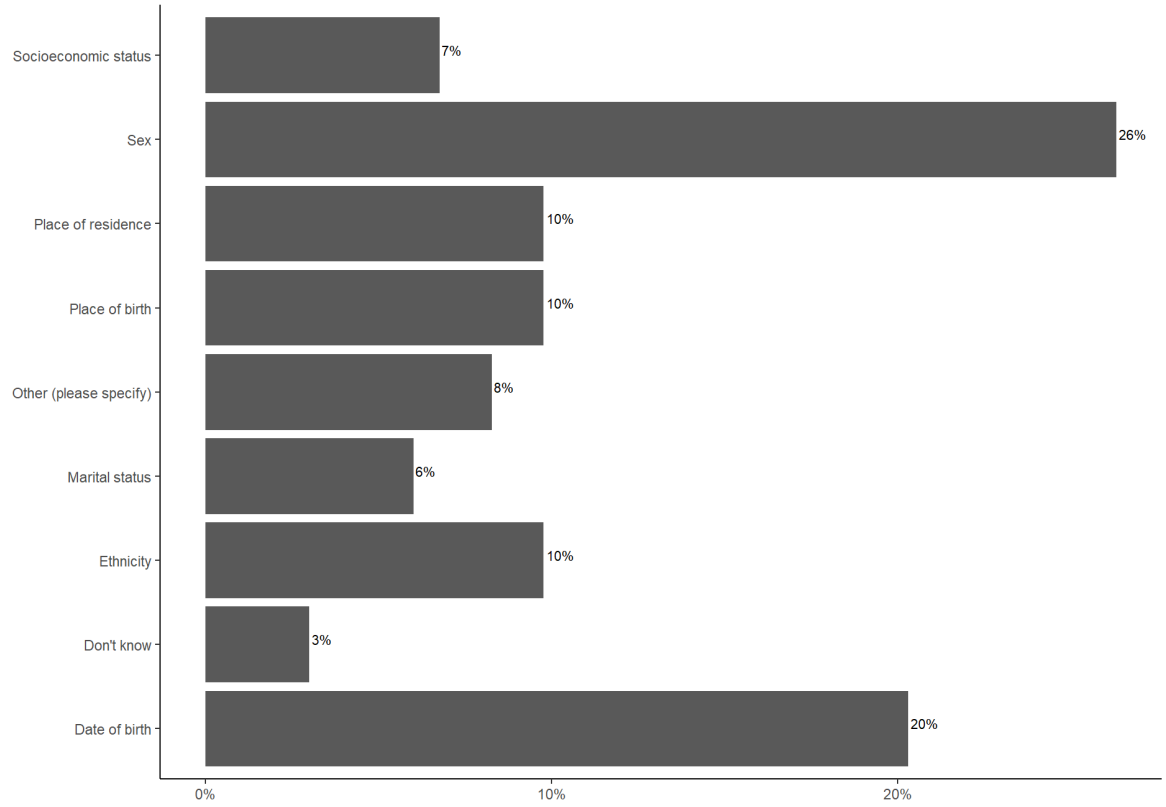
| | Is demographic data being collected |
|---|---|
| | : |
| **Biobank (n=4)** | No: 1 (25.0%)<br>Yes: 3 (75.0%)<br>Missing: 0 (0%) |
| | : |
| **Cohort study (n=20)** | No: 3 (15.0%)<br>Yes: 17 (85.0%)<br>Missing: 0 (0%) |
| | : |
| **Consortium study (n=15)** | No: 6 (40.0%)<br>Yes: 8 (53.3%)<br>Missing: 1 (6.7%) |
| | : |
| **Health administrative data (n=4)** | No: 0 (0%)<br>Yes: 4 (100%)<br>Missing: 0 (0%) |
| | : |
| **Other (please specify below) (n=12)** | No: 6 (50.0%)<br>Yes: 6 (50.0%)<br>Missing: 0 (0%) |
| | : |
| **Registry (n=9)** | No: 1 (11.1%)<br>Yes: 7 (77.8%)<br>Missing: 1 (11.1%) |

Answers divided by purpose of study:

| | Is demographic data being collected |
|---|---|
| | : |
| **Administrative health care data (n=4)** | Yes: 4 (100%)<br>No: 0 (0%)<br>Missing: 0 (0%) |
| | : |
| **Determine potential treatment targets, biomarkers (n=4)** | Yes: 2 (50.0%)<br>No: 2 (50.0%)<br>Missing: 0 (0%) |
| | : |
| **Don't know (n=1)** | Yes: 1 (100%)<br>No: 0 (0%)<br>Missing: 0 (0%) |
| | : |
| **Other (please specify below) (n=3)** | Yes: 3 (100%)<br>No: 0 (0%)<br>Missing: 0 (0%) |
| | : |
| **Patient/cohort classification (n=1)** | Yes: 1 (100%)<br>No: 0 (0%)<br>Missing: 0 (0%) |
| | : |
| **Determine effects of treatment (n=1)** | Yes: 0 (0%)<br>No: 1 (100%)<br>Missing: 0 (0%) |
| **Multiple aims (n=37)** | Yes: 24 (64.9%)<br>No: 13 (35.1%)<br>Missing: 0 (0%) |
| | : |
| **Predict disease onset (n=4)** | Yes: 3 (75.0%)<br>No: 0 (0%)<br>Missing: 1 (25.0%) |
| | : |
| **Overall (n=83)** | Yes: 45 (54.2%)<br>No: 20 (24.1%)<br>Missing: 18 (21.7%) |

## 8. What demographic data is being collected?

| | Overall (n=133) |
|---|---|
| Date of birth | 27 (20.3%) |
| Don't know | 4 (3.0%) |
| Ethnicity | 13 (9.8%) |
| Marital status | 8 (6.0%) |
| Other (please specify) | 11 (8.3%) |
| Place of birth | 13 (9.8%) |
| Place of residence | 13 (9.8%) |
| Sex | 35 (26.3%) |
| Socioeconomic status | 9 (6.8%) |

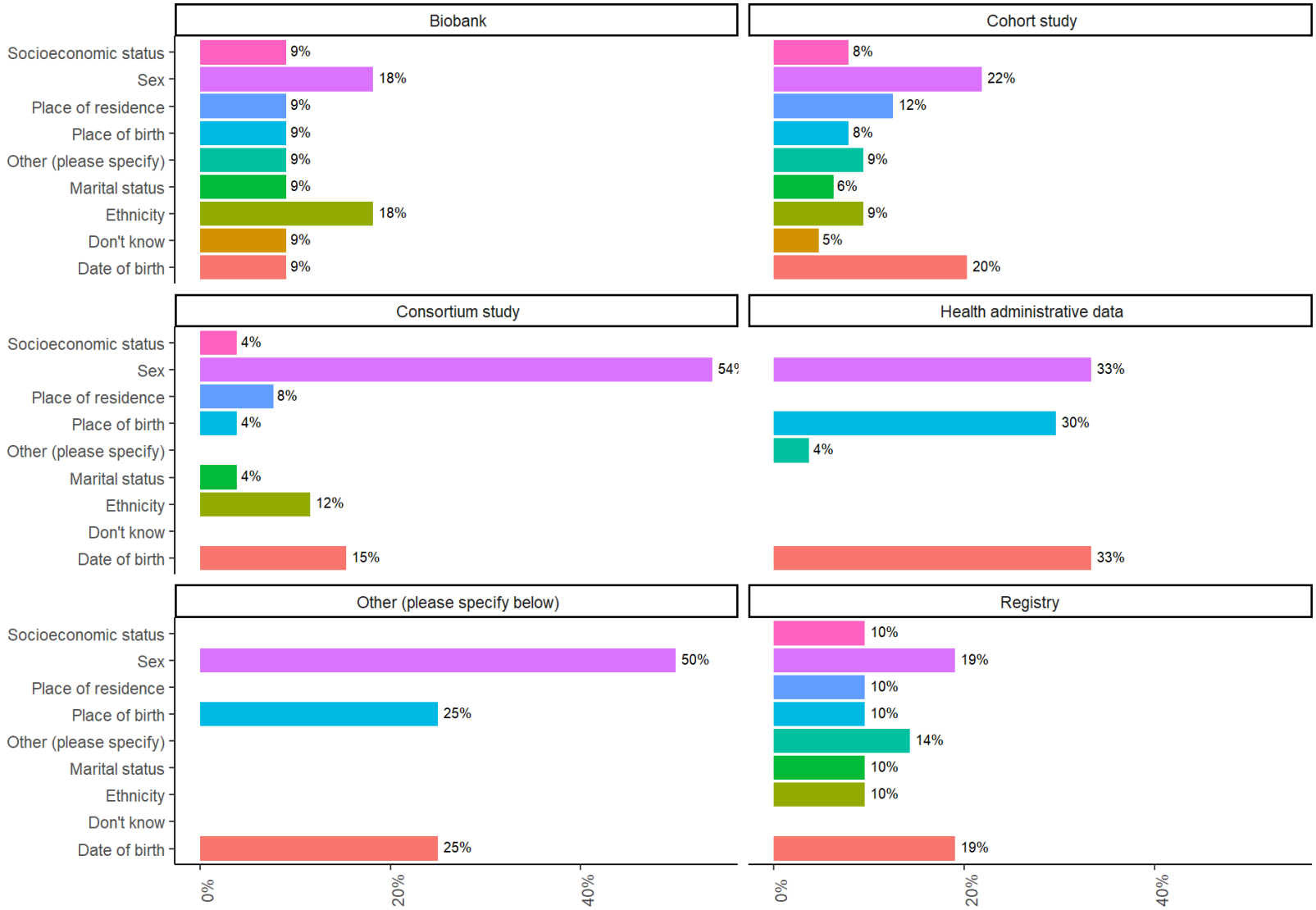


Specifications for those who answered “Other”

| | Overall (n=17) |
|---|---|
| **Other** | |
| age, sex, ethnicity | 1 (5.9%) |
| All clinical and health records | 1 (5.9%) |
| Depends on specific biobank/collection. | 1 (5.9%) |
| for some datasets also other such as birth date, ethnicity, ... | 1 (5.9%) |
| I expect that it varies between cohorts, but includes common risk factors for dementias, including age. | 1 (5.9%) |
| Life style factors, symptoms of disease | 8 (47.1%) |
| Sick leave | 3 (17.6%) |
| We have week of birth, gender and area of residence | 1 (5.9%) |

Answers divided by type of study:

| | Biobank (n=11) | Cohort study (n=64) | Consortium study (n=26) | Health administrative data (n=27) | Other (please specify below) (n=32) | Registry (n=21) | Overall (n=181) |
|---|---|---|---|---|---|---|---|
| **Demographic_data** | | | | | | | |
| Date of birth | 1 (9.1%) | 13 (20.3%) | 4 (15.4%) | 9 (33.3%) | 8 (25.0%) | 4 (19.0%) | 39 (21.5%) |
| Don't know | 1 (9.1%) | 3 (4.7%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 4 (2.2%) |
| Ethnicity | 2 (18.2%) | 6 (9.4%) | 3 (11.5%) | 0 (0%) | 0 (0%) | 2 (9.5%) | 13 (7.2%) |
| Marital status | 1 (9.1%) | 4 (6.2%) | 1 (3.8%) | 0 (0%) | 0 (0%) | 2 (9.5%) | 8 (4.4%) |
| Other (please specify) | 1 (9.1%) | 6 (9.4%) | 0 (0%) | 1 (3.7%) | 0 (0%) | 3 (14.3%) | 11 (6.1%) |
| Place of birth | 1 (9.1%) | 5 (7.8%) | 1 (3.8%) | 8 (29.6%) | 8 (25.0%) | 2 (9.5%) | 25 (13.8%) |
| Place of residence | 1 (9.1%) | 8 (12.5%) | 2 (7.7%) | 0 (0%) | 0 (0%) | 2 (9.5%) | 13 (7.2%) |
| Sex | 2 (18.2%) | 14 (21.9%) | 14 (53.8%) | 9 (33.3%) | 16 (50.0%) | 4 (19.0%) | 59 (32.6%) |
| Socioeconomic status | 1 (9.1%) | 5 (7.8%) | 1 (3.8%) | 0 (0%) | 0 (0%) | 2 (9.5%) | 9 (5.0%) |


Biobank
Socioeconomic status 9%
Sex 18%
Place of residence 9%
Place of birth 9%
Other (please specify) 9%
Marital status 9%
Ethnicity 18%
Don't know 9%
Date of birth 9%
Cohort study
8%
22%
12%
8%
9%
6%
9%
5%
20%
Consortium study
4%
54%
8%
4%
4%
12%
15%
Health administrative data
33%
30%
4%
33%
Other (please specify below)
50%
25%
25%
Registry
10%
19%
10%
10%
14%
10%
10%
19%
0%
20%
40%

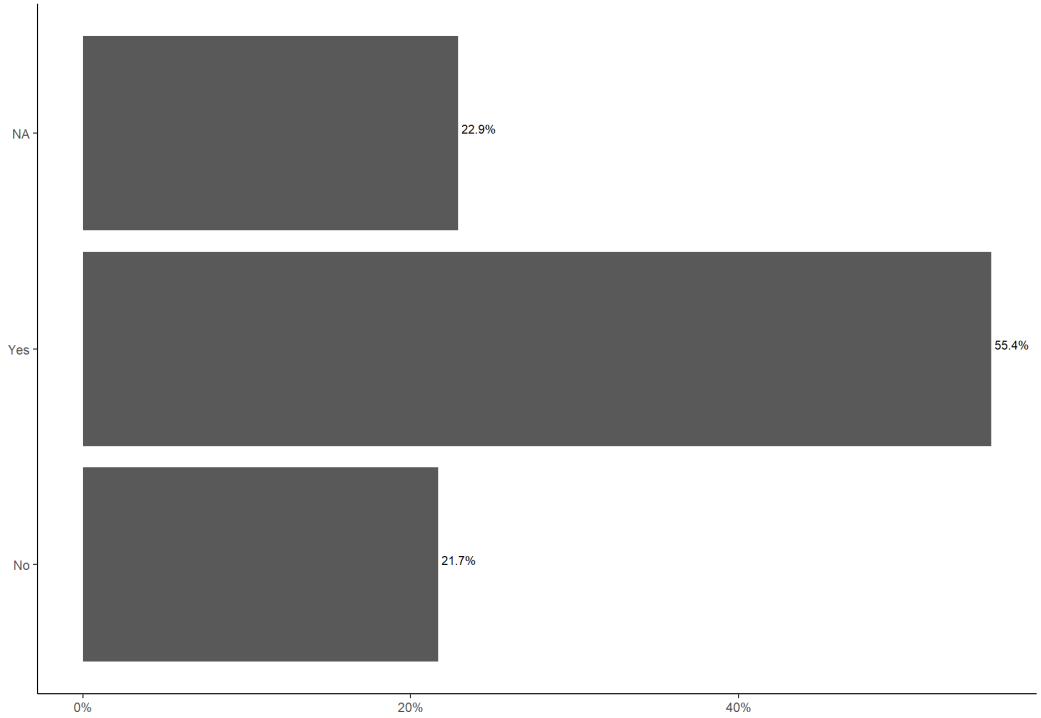
Answers divided by purpose of study:

| | Administrative health care data (n=11) | Determine potential treatment targets, biomarkers (n=5) | Don't know (n=2) | Other (please specify below) (n=15) | Identify disease risk / progression (n=8) | Multiple aims (n=120) | Predict disease onset (n=19) | Overall (n=181) |
|---|---|---|---|---|---|---|---|---|
| **Demographic_data** | | | | | | | | |
| Date of birth | 2 (18.2%) | 1 (20.0%) | 0 (0%) | 3 (20.0%) | 3 (37.5%) | 28 (23.3%) | 2 (10.5%) | 39 (21.5%) |
| Ethnicity | 1 (9.1%) | 1 (20.0%) | 1 (50.0%) | 1 (6.7%) | 1 (12.5%) | 6 (5.0%) | 2 (10.5%) | 13 (7.2%) |
| Marital status | 1 (9.1%) | 0 (0%) | 0 (0%) | 1 (6.7%) | 0 (0%) | 4 (3.3%) | 2 (10.5%) | 8 (4.4%) |
| Other (please specify) | 2 (18.2%) | 1 (20.0%) | 0 (0%) | 2 (13.3%) | 0 (0%) | 4 (3.3%) | 2 (10.5%) | 11 (6.1%) |
| Place of birth | 1 (9.1%) | 0 (0%) | 0 (0%) | 1 (6.7%) | 0 (0%) | 21 (17.5%) | 2 (10.5%) | 25 (13.8%) |
| Place of residence | 1 (9.1%) | 1 (20.0%) | 0 (0%) | 2 (13.3%) | 0 (0%) | 7 (5.8%) | 2 (10.5%) | 13 (7.2%) |
| Sex | 2 (18.2%) | 1 (20.0%) | 1 (50.0%) | 3 (20.0%) | 4 (50.0%) | 45 (37.5%) | 3 (15.8%) | 59 (32.6%) |
| Socioeconomic status | 1 (9.1%) | 0 (0%) | 0 (0%) | 1 (6.7%) | 0 (0%) | 5 (4.2%) | 2 (10.5%) | 9 (5.0%) |
| Don't know | 0 (0%) | 0 (0%) | 0 (0%) | 1 (6.7%) | 0 (0%) | 0 (0%) | 2 (10.5%) | 4 (2.2%) |


Administrative health care data
Determine potential treatment targets, biomarkers
Don't know
Other (please specify below)
Identify disease risk / progression
Multiple aims
Predict disease onset
NA
Socioeconomic status
Sex
Place of residence
Place of birth
Other (please specify)
Marital status
Ethnicity
Don't know
Date of birth
0%
25%
50%
75%
100%

## 9. Is biological data being collected?

| | Overall (n=83) |
|---|---|
| **Is biological data being collected** | |
| No | 18 (21.7%) |
| Yes | 46 (55.4%) |
| Missing | 19 (22.9%) |



Answers divided by type of study:

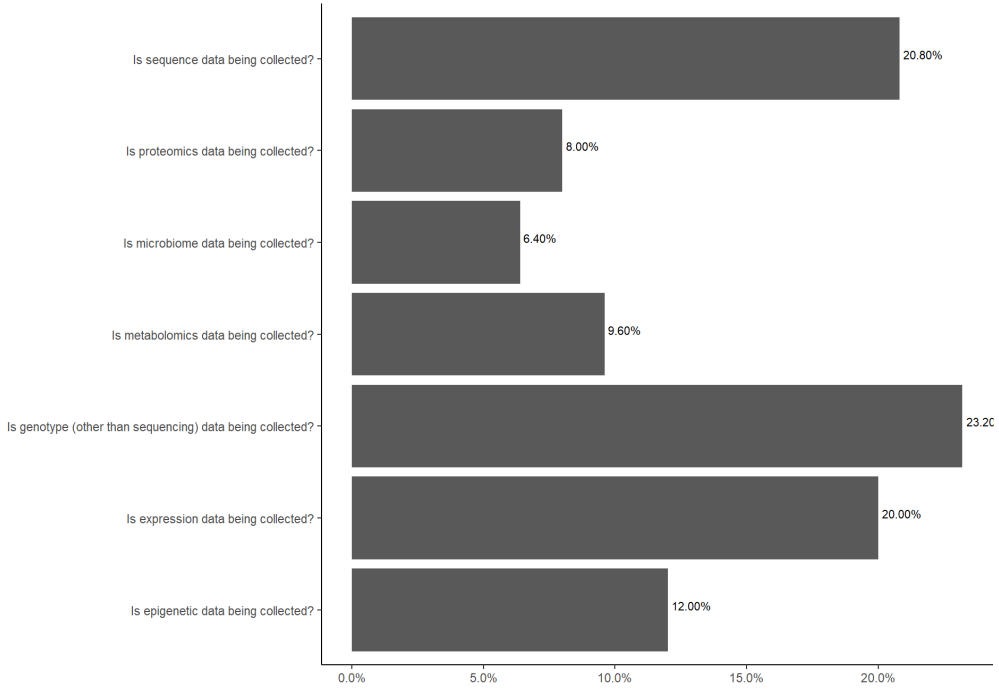
| | Is biological data being collected |
|---|---|
| | : |
| **Biobank (n=4)** | Yes: 4 (100%)<br>No: 0 (0%)<br>Missing: 0 (0%) |
| | : |
| **Cohort study (n=20)** | Yes: 18 (90.0%)<br>No: 0 (0%)<br>Missing: 2 (10.0%) |
| | : |
| **Consortium study (n=15)** | Yes: 13 (86.7%)<br>No: 2 (13.3%)<br>Missing: 0 (0%) |
| | : |
| **Health administrative data (n=4)** | Yes: 0 (0%)<br>No: 4 (100%)<br>Missing: 0 (0%) |
| | : |
| **Other (please specify below) (n=12)** | Yes: 6 (50.0%)<br>No: 5 (41.7%)<br>Missing: 1 (8.3%) |
| | : |
| **Registry (n=9)** | Yes: 3 (33.3%)<br>No: 6 (66.7%)<br>Missing: 0 (0%) |
| | : |
| **Overall (n=83)** | Yes: 46 (55.4%)<br>No: 18 (21.7%)<br>Missing: 19 (22.9%) |

Answers divided by purpose of study:

| | Is biological data being collected |
|---|---|
| | : |
| **Administrative health care data (n=4)** | No: 2 (50.0%)<br>Yes: 2 (50.0%)<br>Missing: 0 (0%) |
| | : |
| **Determine potential treatment targets, biomarkers (n=4)** | No: 0 (0%)<br>Yes: 4 (100%)<br>Missing: 0 (0%) |
| | : |
| **Don't know (n=1)** | No: 0 (0%)<br>Yes: 1 (100%)<br>Missing: 0 (0%) |
| | : |
| **Other (please specify below) (n=3)** | No: 1 (33.3%)<br>Yes: 2 (66.7%)<br>Missing: 0 (0%) |
| | : |
| **Patient/cohort classification (n=1)** | No: 1 (100%)<br>Yes: 0 (0%)<br>Missing: 0 (0%) |
| | : |
| **Determine effects of treatment (n=1)** | No: 0 (0%)<br>Yes: 1 (100%)<br>Missing: 0 (0%) |
| **Identify disease risk / progression (n=6)** | No: 1 (16.7%)<br>Yes: 5 (83.3%)<br>Missing: 0 (0%) |
| | : |
| **Multiple aims (n=37)** | No: 12 (32.4%)<br>Yes: 25 (67.6%)<br>Missing: 0 (0%) |
| | : |
| **Predict disease onset (n=4)** | No: 0 (0%)<br>Yes: 3 (75.0%)<br>Missing: 1 (25.0%) |
| | : |
| **Overall (n=83)** | No: 18 (21.7%)<br>Yes: 46 (55.4%)<br>Missing: 19 (22.9%) |

## 10. Is biological data being collected? What type?

| | Overall (n=125) |
|---|---|
| **When collected, what type of biological data is being collected?** | |
| Is epigenetic data being collected? | 15 (12.0%) |
| Is expression data being collected? | 25 (20.0%) |
| Is genotype (other than sequencing) data being collected? | 29 (23.2%) |
| Is metabolomics data being collected? | 12 (9.6%) |
| Is microbiome data being collected? | 8 (6.4%) |
| Is proteomics data being collected? | 10 (8.0%) |
| Is sequence data being collected? | 26 (20.8%) |



Answers divided by type of study:

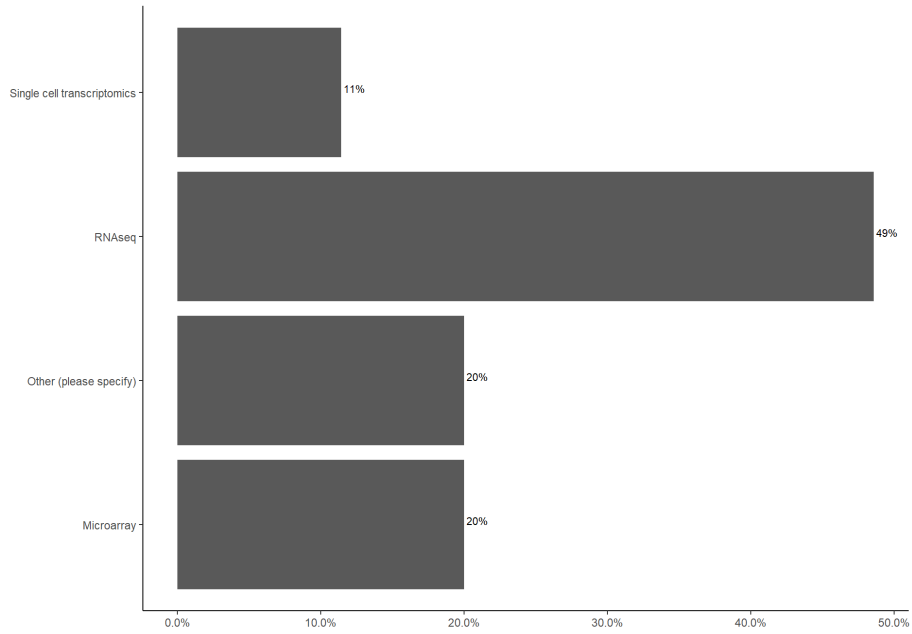
| | Biobank (n=6) | Cohort study (n=55) | Consortium study (n=68) | Other (please specify below) (n=44) | Registry (n=8) | Overall (n=185) |
|---|---|---|---|---|---|---|
| **Biological_data** | | | | | | |
| Is genotype (other than sequencing) data being collected? | 4 (66.7%) | 10 (18.2%) | 15 (22.1%) | 9 (20.5%) | 2 (25.0%) | 41 (22.2%) |
| Is microbiome data being collected? | 1 (16.7%) | 4 (7.3%) | 2 (2.9%) | 0 (0%) | 1 (12.5%) | 8 (4.3%) |
| Is sequence data being collected? | 1 (16.7%) | 6 (10.9%) | 22 (32.4%) | 17 (38.6%) | 2 (25.0%) | 50 (27.0%) |
| Is epigenetic data being collected? | 0 (0%) | 7 (12.7%) | 11 (16.2%) | 9 (20.5%) | 0 (0%) | 27 (14.6%) |
| Is expression data being collected? | 0 (0%) | 13 (23.6%) | 13 (19.1%) | 9 (20.5%) | 1 (12.5%) | 37 (20.0%) |
| Is metabolomics data being collected? | 0 (0%) | 8 (14.5%) | 3 (4.4%) | 0 (0%) | 1 (12.5%) | 12 (6.5%) |
| Is proteomics data being collected? | 0 (0%) | 7 (12.7%) | 2 (2.9%) | 0 (0%) | 1 (12.5%) | 10 (5.4%) |

Answers divided by purpose of study:

| | Administrative health care data (n=2) | Determine potential treatment targets, biomarkers (n=9) | Don't know (n=1) | Other (please specify below) (n=7) | Determine effects of treatment (n=1) | Identify disease risk / progression (n=15) | Multiple aims (n=132) | Predict disease onset (n=11) | Overall (n=185) |
|---|---|---|---|---|---|---|---|---|---|
| **Biological_data** | | | | | | | | | |
| Is genotype (other than sequencing) data being collected? | 1 (50.0%) | 2 (22.2%) | 1 (100%) | 2 (28.6%) | 0 (0%) | 3 (20.0%) | 27 (20.5%) | 3 (27.3%) | 41 (22.2%) |
| Is sequence data being collected? | 1 (50.0%) | 1 (11.1%) | 0 (0%) | 1 (14.3%) | 0 (0%) | 4 (26.7%) | 42 (31.8%) | 1 (9.1%) | 50 (27.0%) |
| Is epigenetic data being collected? | 0 (0%) | 1 (11.1%) | 0 (0%) | 2 (28.6%) | 0 (0%) | 1 (6.7%) | 20 (15.2%) | 2 (18.2%) | 27 (14.6%) |
| Is expression data being collected? | 0 (0%) | 3 (33.3%) | 0 (0%) | 1 (14.3%) | 1 (100%) | 3 (20.0%) | 25 (18.9%) | 2 (18.2%) | 37 (20.0%) |
| Is metabolomics data being collected? | 0 (0%) | 1 (11.1%) | 0 (0%) | 1 (14.3%) | 0 (0%) | 1 (6.7%) | 7 (5.3%) | 1 (9.1%) | 12 (6.5%) |
| Is microbiome data being collected? | 0 (0%) | 1 (11.1%) | 0 (0%) | 0 (0%) | 0 (0%) | 1 (6.7%) | 5 (3.8%) | 1 (9.1%) | 8 (4.3%) |
| Is proteomics data being collected? | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 2 (13.3%) | 6 (4.5%) | 1 (9.1%) | 10 (5.4%) |

### 11. What expression data is being collected?

| | Overall (n=35) |
|---|---|
| Microarray | 7 (20.0%) |
| Other (please specify) | 7 (20.0%) |
| RNAseq | 17 (48.6%) |
| Single cell transcriptomics | 4 (11.4%) |

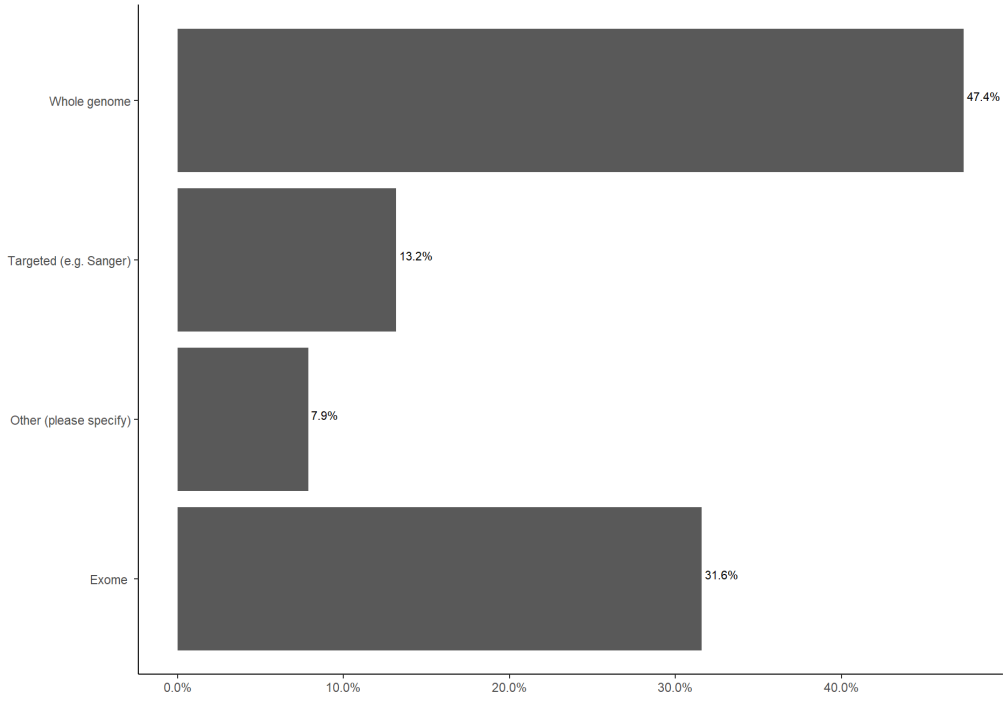



Specifications of those answering other:

| | Overall (n=5) |
|---|---|
| **Other** | |
| Affynetrix arrays | 1 (20.0%) |
| Depends on specific biobank/collection. | 1 (20.0%) |
| mRNA, miRNA, proteome/peptidome | 1 (20.0%) |
| Not defined yet | 1 (20.0%) |
| serum miRNA | 1 (20.0%) |

**12. What sequence data is being collected?**

| | Overall (n=38) |
|---|---|
| Exome | 12 (31.6%) |
| Other (please specify) | 3 (7.9%) |
| Targeted (e.g. Sanger) | 5 (13.2%) |
| Whole genome | 18 (47.4%) |

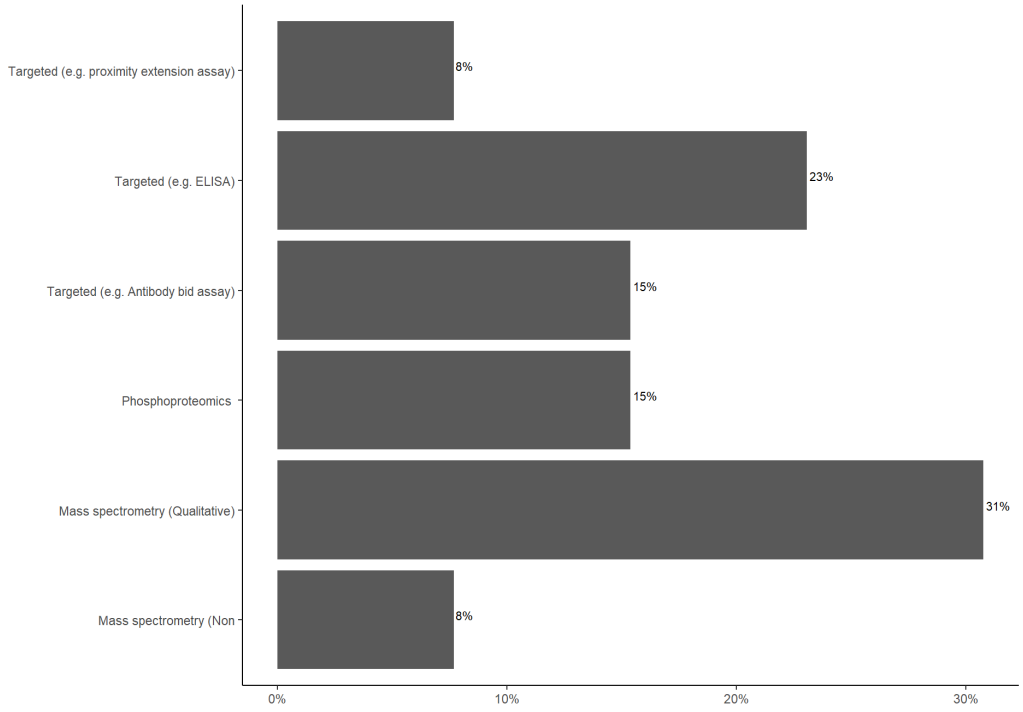



Specification of those answering other:

| | Overall (n=1) |
|---|---|
| **Other** | |
| Depends on specific biobank | 1 (100%) |

**13. What proteomics data is being collected?**

| | Overall (n=13) |
|---|---|
| Mass spectrometry (Non | 1 (7.7%) |
| Mass spectrometry (Qualitative) | 4 (30.8%) |
| Phosphoproteomics | 2 (15.4%) |
| Targeted (e.g. Antibody bid assay) | 2 (15.4%) |
| Targeted (e.g. ELISA) | 3 (23.1%) |
| Targeted (e.g. proximity extension assay) | 1 (7.7%) |

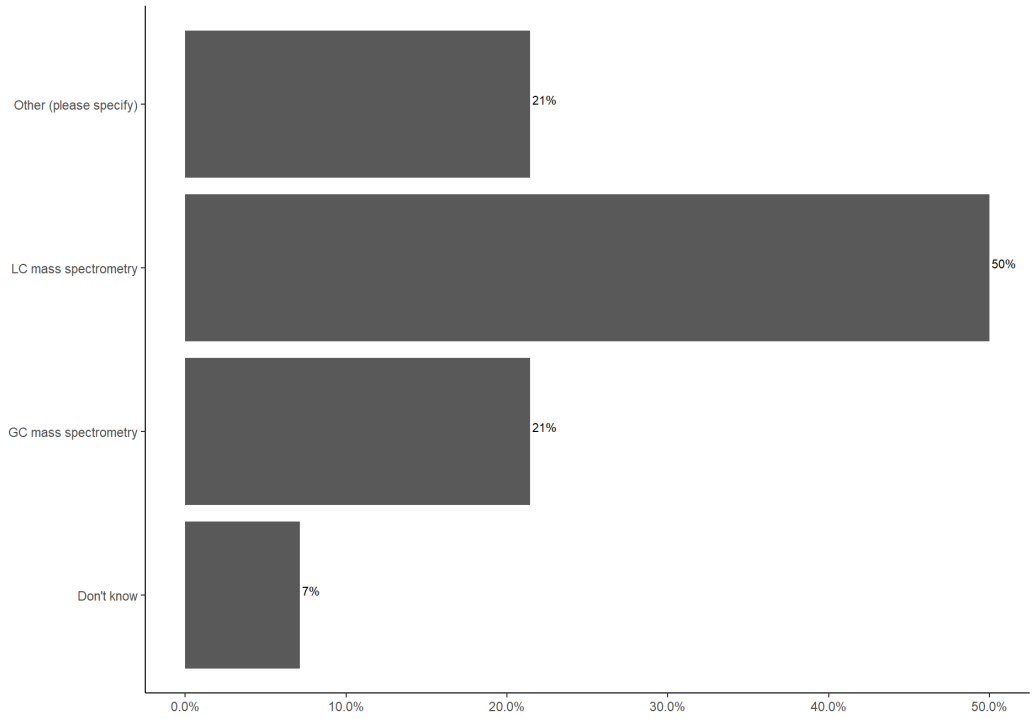



Specification of those answering other:

| | Overall (n=2) |
|---|---|
| **Other** | |
| Depends on specific biobank | 1 (50.0%) |
| e.g. BDNF | 1 (50.0%) |

**14. What metabolomics data is being collected?**

| | Overall (n=14) |
|---|---|
| Don't know | 1 (7.1%) |
| GCmass spectrometry | 3 (21.4%) |
| LCmass spectrometry | 7 (50.0%) |
| Other (please specify) | 3 (21.4%) |

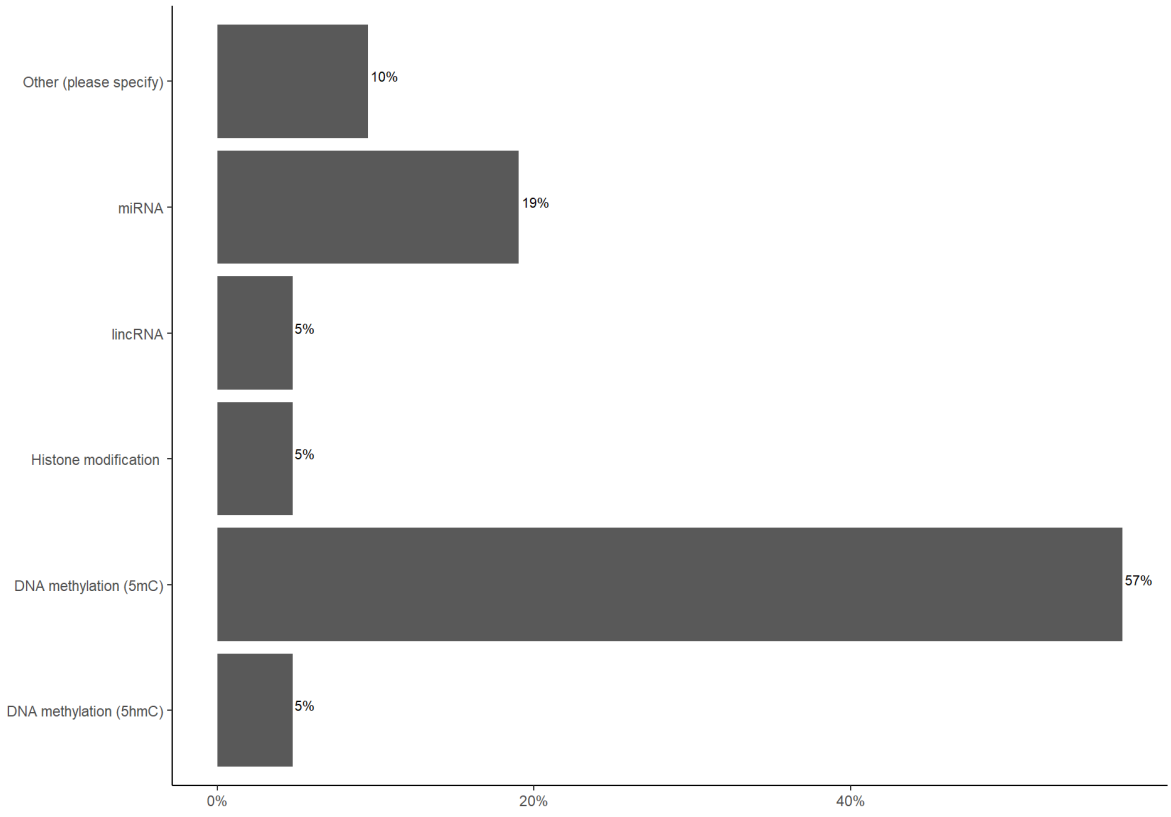



Specification of those answering other:

| | Overall (n=3) |
|---|---|
| **Other** | |
| Depends on specific biobank | 1 (33.3%) |
| e.g. Lipidomics | 1 (33.3%) |
| HPLC, ELISA | 1 (33.3%) |

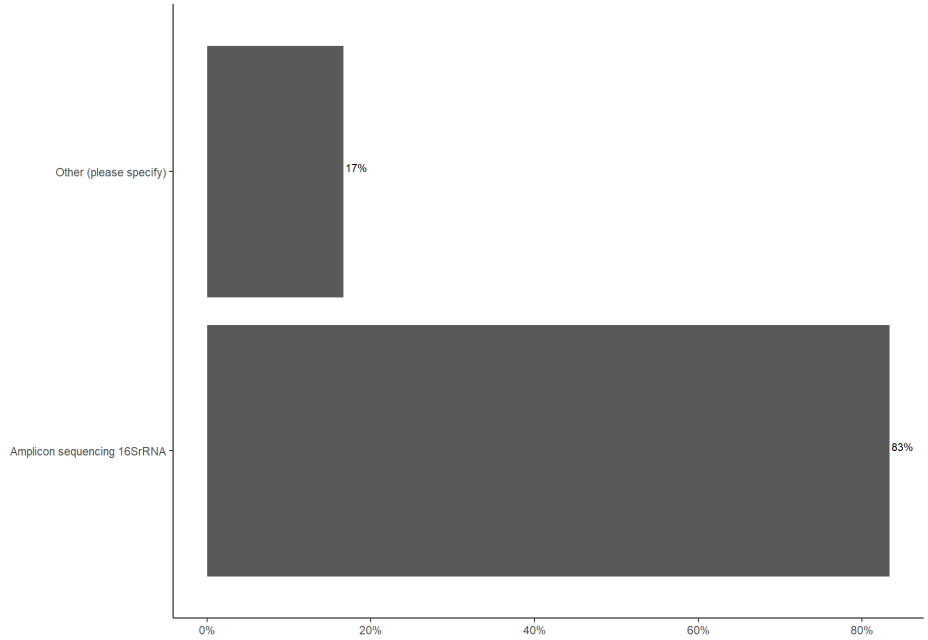
**15. What epigenetic data is being collected?**

| | Overall (n=21) |
|---|---|
| DNA methylation (5hmC) | 1 (4.8%) |
| DNA methylation (5mC) | 12 (57.1%) |
| Histone modification | 1 (4.8%) |
| lincRNA | 1 (4.8%) |
| miRNA | 4 (19.0%) |
| Other (please specify) | 2 (9.5%) |



**16. What microbiome data is being collected?**

| | Overall (n=6) |
|---|---|
| Amplicon sequencing 16SrRNA | 5 (83.3%) |
| Other (please specify) | 1 (16.7%) |

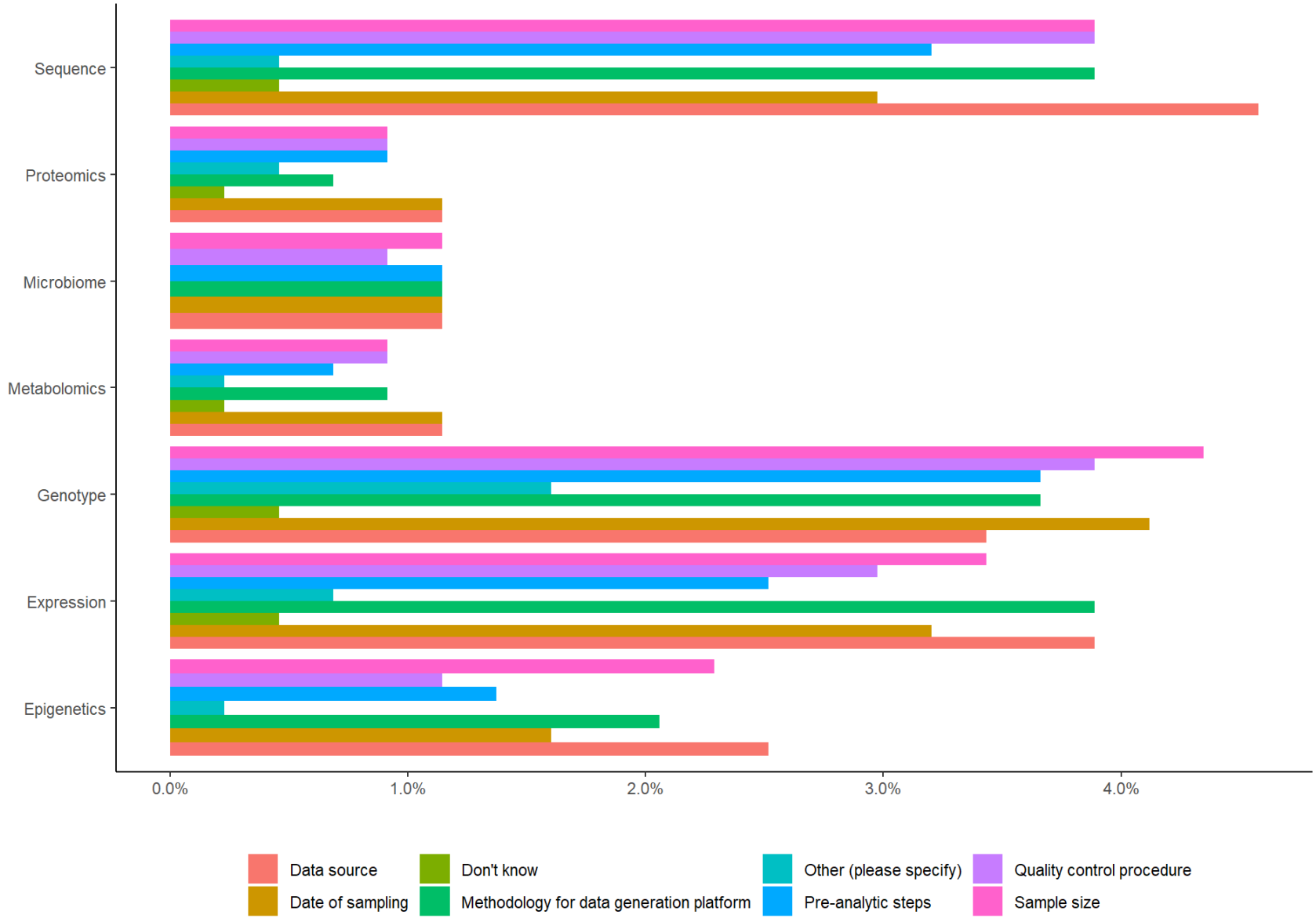


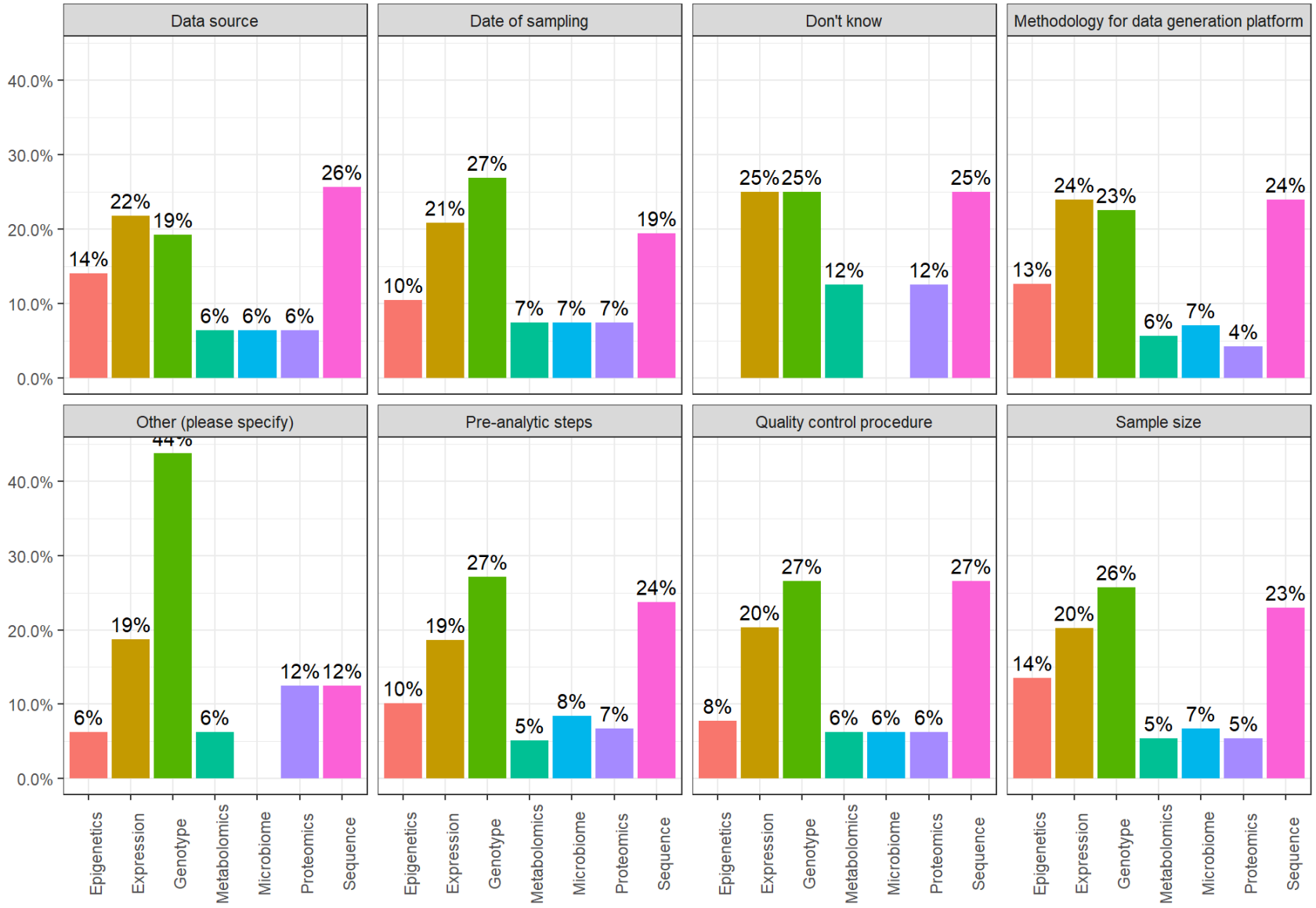

Specifications for those answering other:

| | Overall (n=2) |
|---|---|
| **Other** | |
| Depends on specific biobank/collection. | 1 (50.0%) |
| Samples for microbiome is collected from nasal swaps | 1 (50.0%) |

**17. What type of data is being collected? Divided by type of profiling?**

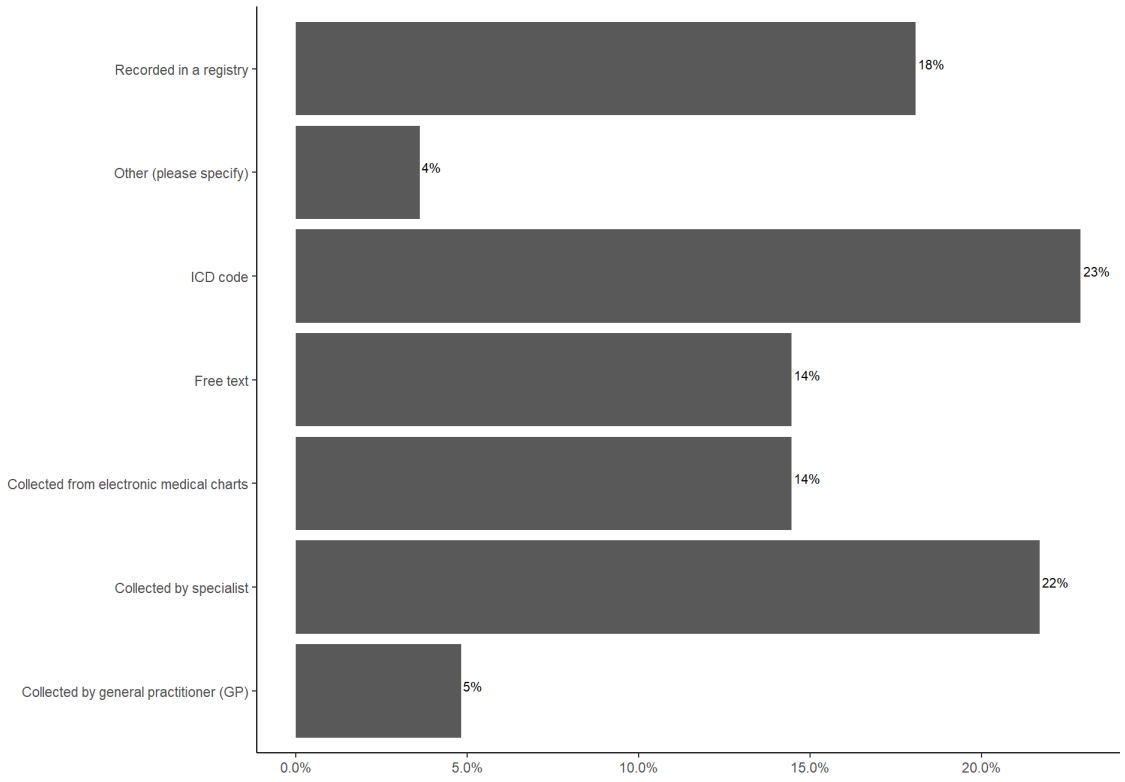
| | Epigenetics (n=49) | Expression (n=92) | Genotype (n=110) | Metabolomics (n=27) | Microbiome (n=29) | Proteomics (n=28) | Sequence (n=102) | Overall (n=437) |
|---|---|---|---|---|---|---|---|---|
| Data source | 11 (22.4%) | 17 (18.5%) | 15 (13.6%) | 5 (18.5%) | 5 (17.2%) | 5 (17.9%) | 20 (19.6%) | 78 (17.8%) |
| Date of sampling | 7 (14.3%) | 14 (15.2%) | 18 (16.4%) | 5 (18.5%) | 5 (17.2%) | 5 (17.9%) | 13 (12.7%) | 67 (15.3%) |
| Methodology for data generation platform | 9 (18.4%) | 17 (18.5%) | 16 (14.5%) | 4 (14.8%) | 5 (17.2%) | 3 (10.7%) | 17 (16.7%) | 71 (16.2%) |
| Other (please specify) | 1 (2.0%) | 3 (3.3%) | 7 (6.4%) | 1 (3.7%) | 0 (0%) | 2 (7.1%) | 2 (2.0%) | 16 (3.7%) |
| Pre-analytic steps | 6 (12.2%) | 11 (12.0%) | 16 (14.5%) | 3 (11.1%) | 5 (17.2%) | 4 (14.3%) | 14 (13.7%) | 59 (13.5%) |
| Quality control procedure | 5 (10.2%) | 13 (14.1%) | 17 (15.5%) | 4 (14.8%) | 4 (13.8%) | 4 (14.3%) | 17 (16.7%) | 64 (14.6%) |
| Sample size | 10 (20.4%) | 15 (16.3%) | 19 (17.3%) | 4 (14.8%) | 5 (17.2%) | 4 (14.3%) | 17 (16.7%) | 74 (16.9%) |
| Don't know | 0 (0%) | 2 (2.2%) | 2 (1.8%) | 1 (3.7%) | 0 (0%) | 1 (3.6%) | 2 (2.0%) | 8 (1.8%) |


Sequence
Proteomics
Microbiome
Metabolomics
Genotype
Expression
Epigenetics
0.0%
1.0%
2.0%
3.0%
4.0%
Data source
Date of sampling
Don't know
Methodology for data generation platform
Other (please specify)
Pre-analytic steps
Quality control procedure
Sample size


Data source
Date of sampling
Don't know
Methodology for data generation platform
Other (please specify)
Pre-analytic steps
Quality control procedure
Sample size
40.0%
30.0%
20.0%
10.0%
0.0%
14% 22% 19% 6% 6% 6% 26%
10% 21% 27% 7% 7% 7% 19%
25% 25% 12% 12% 25%
13% 24% 23% 6% 7% 4% 24%
6% 19% 44% 6% 12% 12%
10% 19% 27% 5% 8% 7% 24%
8% 20% 27% 6% 6% 6% 27%
14% 20% 26% 5% 7% 5% 23%
Epigenetics
Expression
Genotype
Metabolomics
Microbiome
Proteomics
Sequence

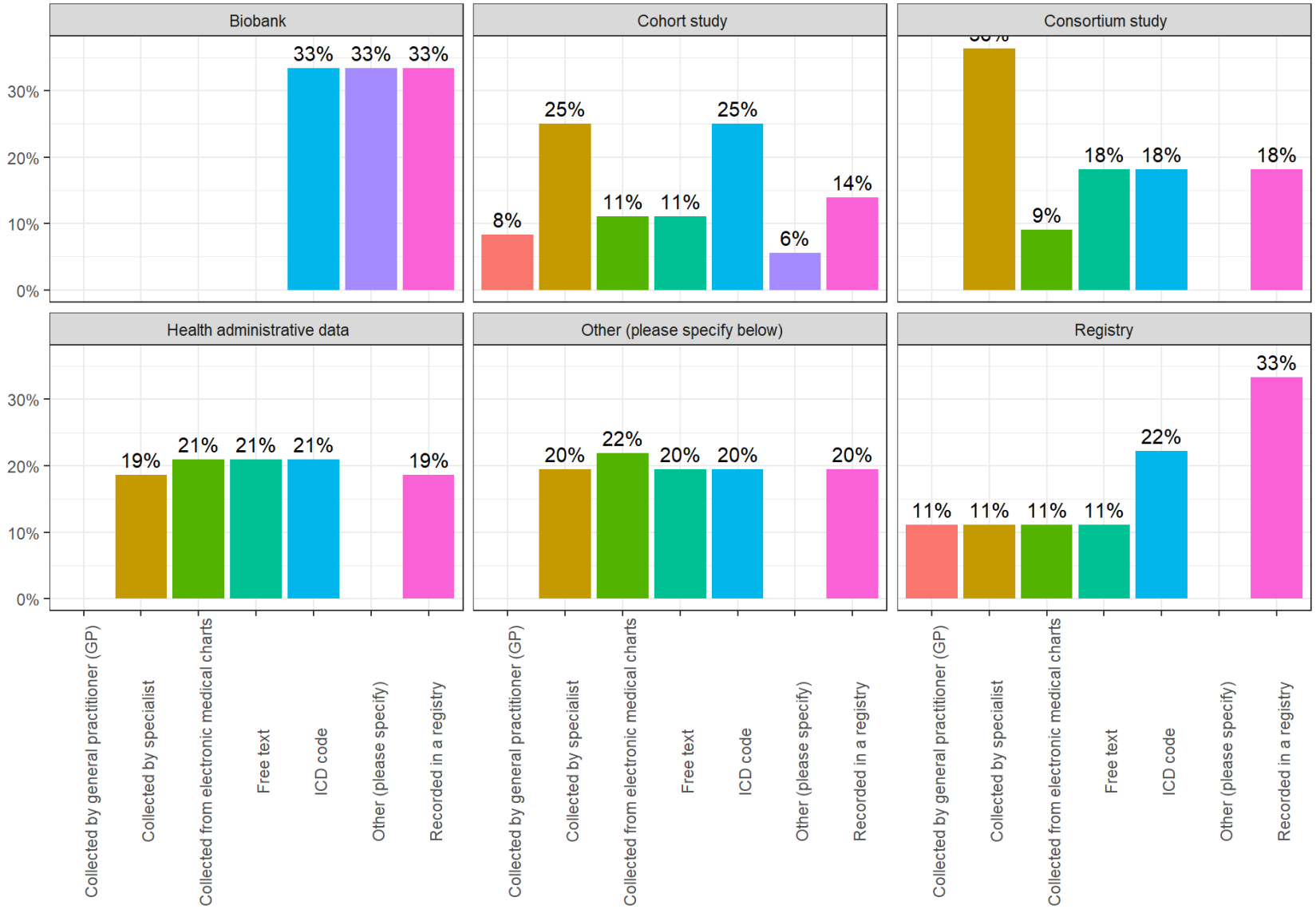
**18. How is the comorbidity data being collected and from which source?**

| | Overall (n=83) |
|---|---|
| **How is the comorbidity data being collected?** | |
| Collected by general practitioner (GP) | 4 (4.8%) |
| Collected by specialist | 18 (21.7%) |
| Collected from electronic medical charts | 12 (14.5%) |
| Free text | 12 (14.5%) |
| ICD code | 19 (22.9%) |
| Other (please specify) | 3 (3.6%) |
| Recorded in a registry | 15 (18.1%) |



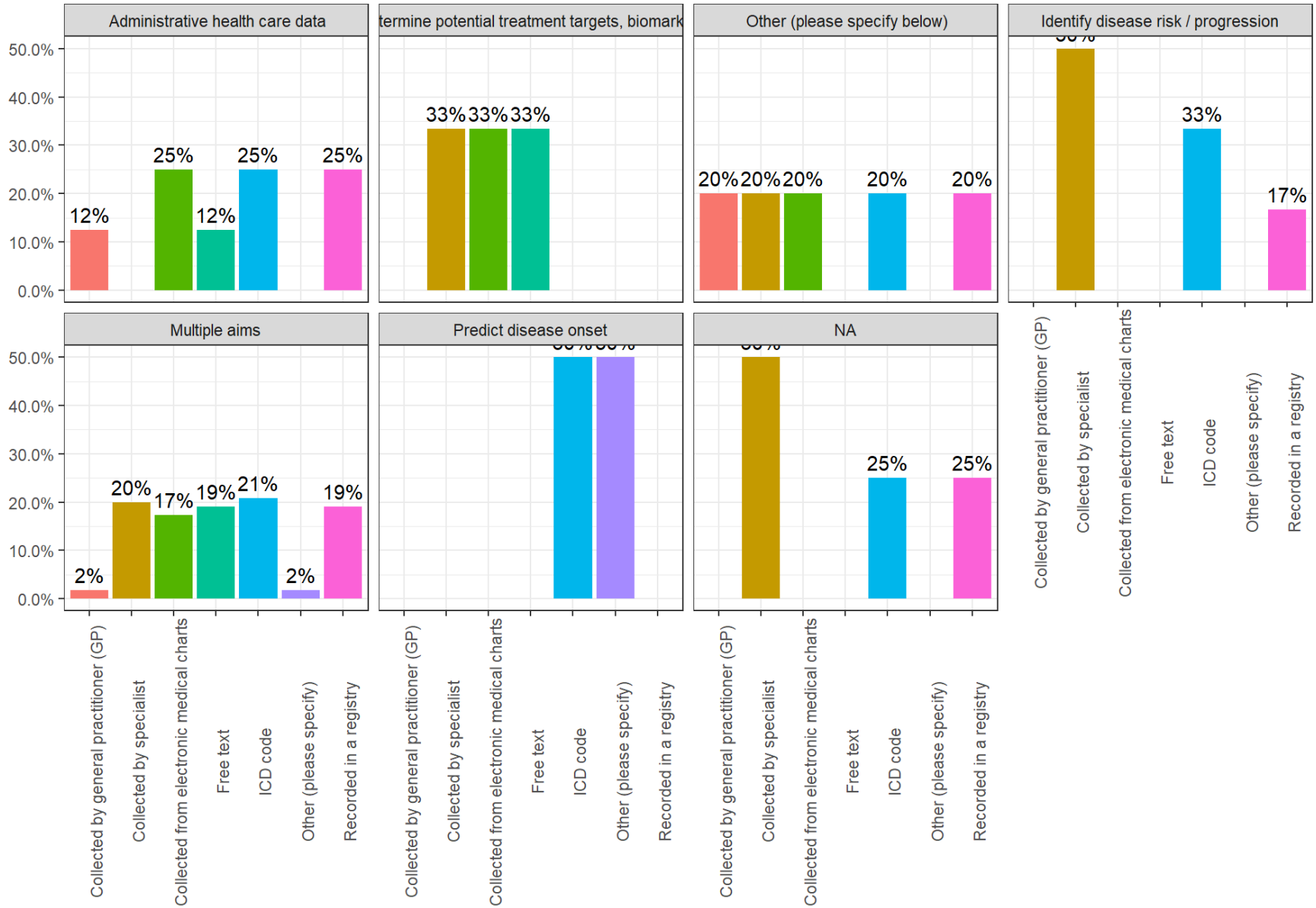
Answerds divided by type of study

| | Biobank (n=3) | Cohort study (n=36) | Consortium study (n=11) | Health administrative data (n=43) | Other (please specify below) (n=41) | Registry (n=9) | Overall (n=143) |
|---|---|---|---|---|---|---|---|
| **Comorbidity_data** | | | | | | | |
| ICD code | 1 (33.3%) | 9 (25.0%) | 2 (18.2%) | 9 (20.9%) | 8 (19.5%) | 2 (22.2%) | 31 (21.7%) |
| Other (please specify) | 1 (33.3%) | 2 (5.6%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 3 (2.1%) |
| Recorded in a registry | 1 (33.3%) | 5 (13.9%) | 2 (18.2%) | 8 (18.6%) | 8 (19.5%) | 3 (33.3%) | 27 (18.9%) |
| Collected by general practitioner (GP) | 0 (0%) | 3 (8.3%) | 0 (0%) | 0 (0%) | 0 (0%) | 1 (11.1%) | 4 (2.8%) |
| Collected by specialist | 0 (0%) | 9 (25.0%) | 4 (36.4%) | 8 (18.6%) | 8 (19.5%) | 1 (11.1%) | 30 (21.0%) |
| Collected from electronic medical charts | 0 (0%) | 4 (11.1%) | 1 (9.1%) | 9 (20.9%) | 9 (22.0%) | 1 (11.1%) | 24 (16.8%) |
| Free text | 0 (0%) | 4 (11.1%) | 2 (18.2%) | 9 (20.9%) | 8 (19.5%) | 1 (11.1%) | 24 (16.8%) |


Biobank
33%
33%
33%
Cohort study
8%
25%
11%
11%
25%
6%
14%
Consortium study
9%
18%
18%
18%
Health administrative data
19%
21%
21%
21%
19%
Other (please specify below)
20%
22%
20%
20%
20%
Registry
11%
11%
11%
11%
22%
33%
30%
20%
10%
0%
Collected by general practitioner (GP)
Collected by specialist
Collected from electronic medical charts
Free text
ICD code
Other (please specify)
Recorded in a registry

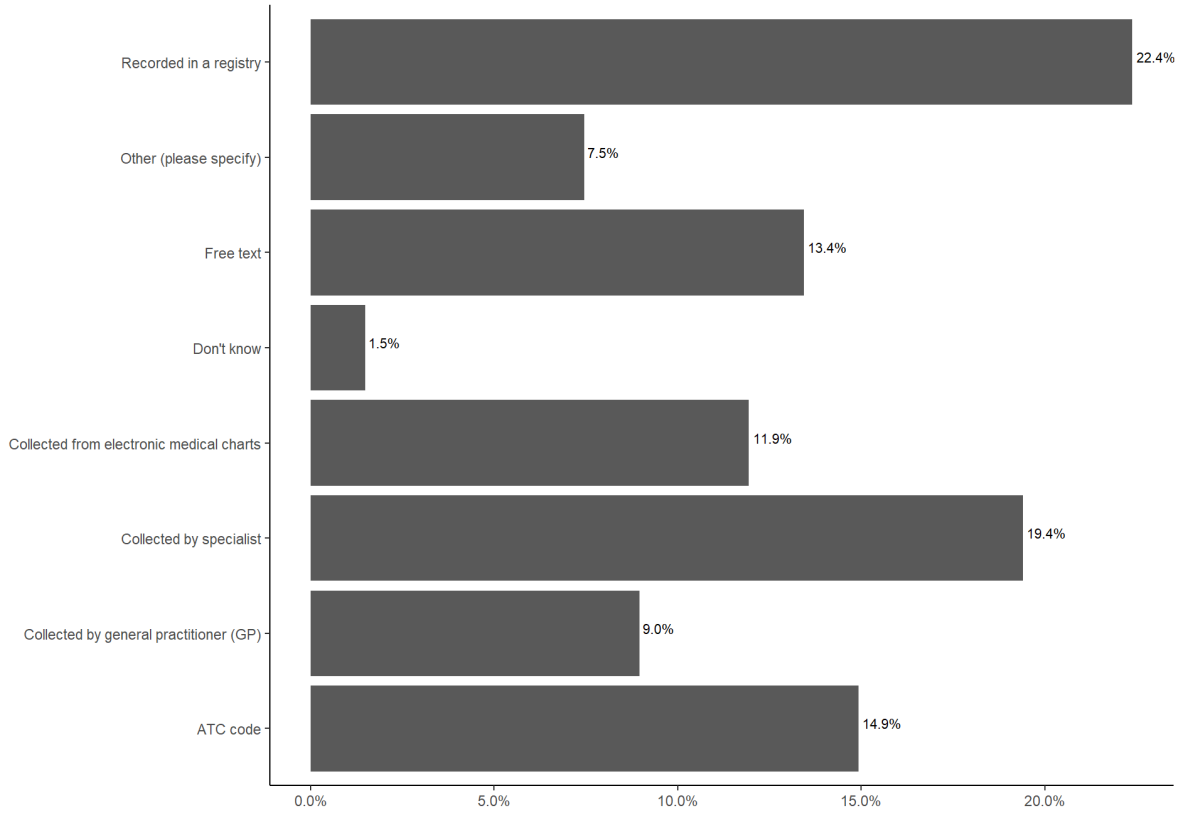
Answers divided by purpose of study:

| | Administrative health care data (n=8) | Determine potential treatment targets, biomarkers (n=3) | Other (please specify below) (n=5) | Identify disease risk / progression (n=6) | Multiple aims (n=115) | Predict disease onset (n=2) | Overall (n=143) |
|---|---|---|---|---|---|---|---|
| **Comorbidity_data** | | | | | | | |
| Collected by general practitioner (GP) | 1 (12.5%) | 0 (0%) | 1 (20.0%) | 0 (0%) | 2 (1.7%) | 0 (0%) | 4 (2.8%) |
| Collected from electronic medical charts | 2 (25.0%) | 1 (33.3%) | 1 (20.0%) | 0 (0%) | 20 (17.4%) | 0 (0%) | 24 (16.8%) |
| Free text | 1 (12.5%) | 1 (33.3%) | 0 (0%) | 0 (0%) | 22 (19.1%) | 0 (0%) | 24 (16.8%) |
| ICD code | 2 (25.0%) | 0 (0%) | 1 (20.0%) | 2 (33.3%) | 24 (20.9%) | 1 (50.0%) | 31 (21.7%) |
| Recorded in a registry | 2 (25.0%) | 0 (0%) | 1 (20.0%) | 1 (16.7%) | 22 (19.1%) | 0 (0%) | 27 (18.9%) |
| Collected by specialist | 0 (0%) | 1 (33.3%) | 1 (20.0%) | 3 (50.0%) | 23 (20.0%) | 0 (0%) | 30 (21.0%) |
| Other (please specify) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 2 (1.7%) | 1 (50.0%) | 3 (2.1%) |


Administrative health care data
termine potential treatment targets, biomark
Other (please specify below)
Identify disease risk / progression
Multiple aims
Predict disease onset
NA
50.0%
40.0%
30.0%
20.0%
10.0%
0.0%
12%
25%
12%
25%
25%
33%
33%
33%
20%
20%
20%
20%
20%
33%
17%
2%
20%
17%
19%
21%
2%
19%
25%
25%
Collected by general practitioner (GP)
Collected by specialist
Collected from electronic medical charts
Free text
ICD code
Other (please specify)
Recorded in a registry

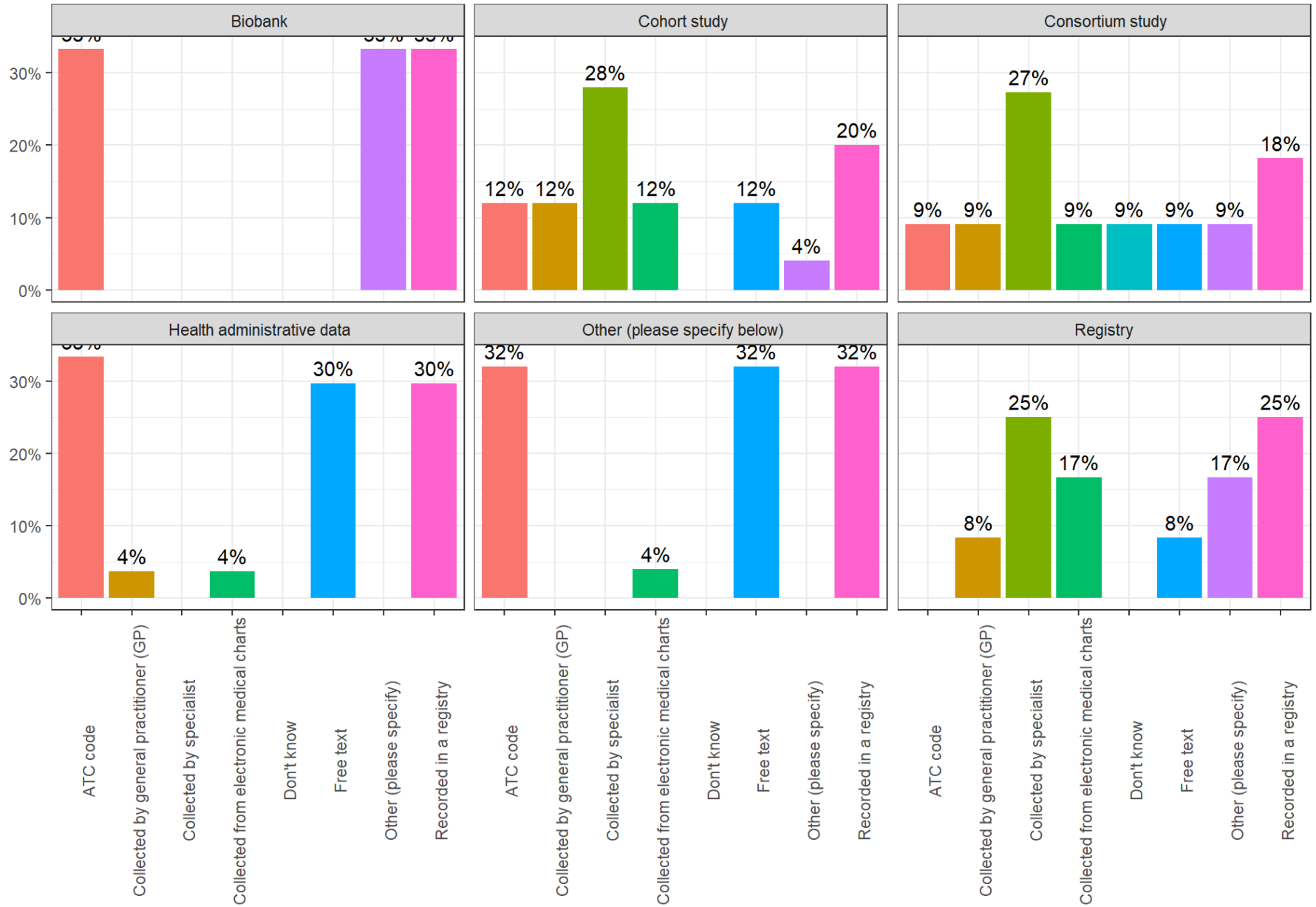
**19. How is the medication data being collected and from which source?**

| | Overall (n=67) |
|---|---|
| **How is the medications data being collected?** | |
| ATC code | 10 (14.9%) |
| Collected by general practitioner (GP) | 6 (9.0%) |
| Collected by specialist | 13 (19.4%) |
| Collected from electronic medical charts | 8 (11.9%) |
| Don't know | 1 (1.5%) |
| Free text | 9 (13.4%) |
| Other (please specify) | 5 (7.5%) |
| Recorded in a registry | 15 (22.4%) |



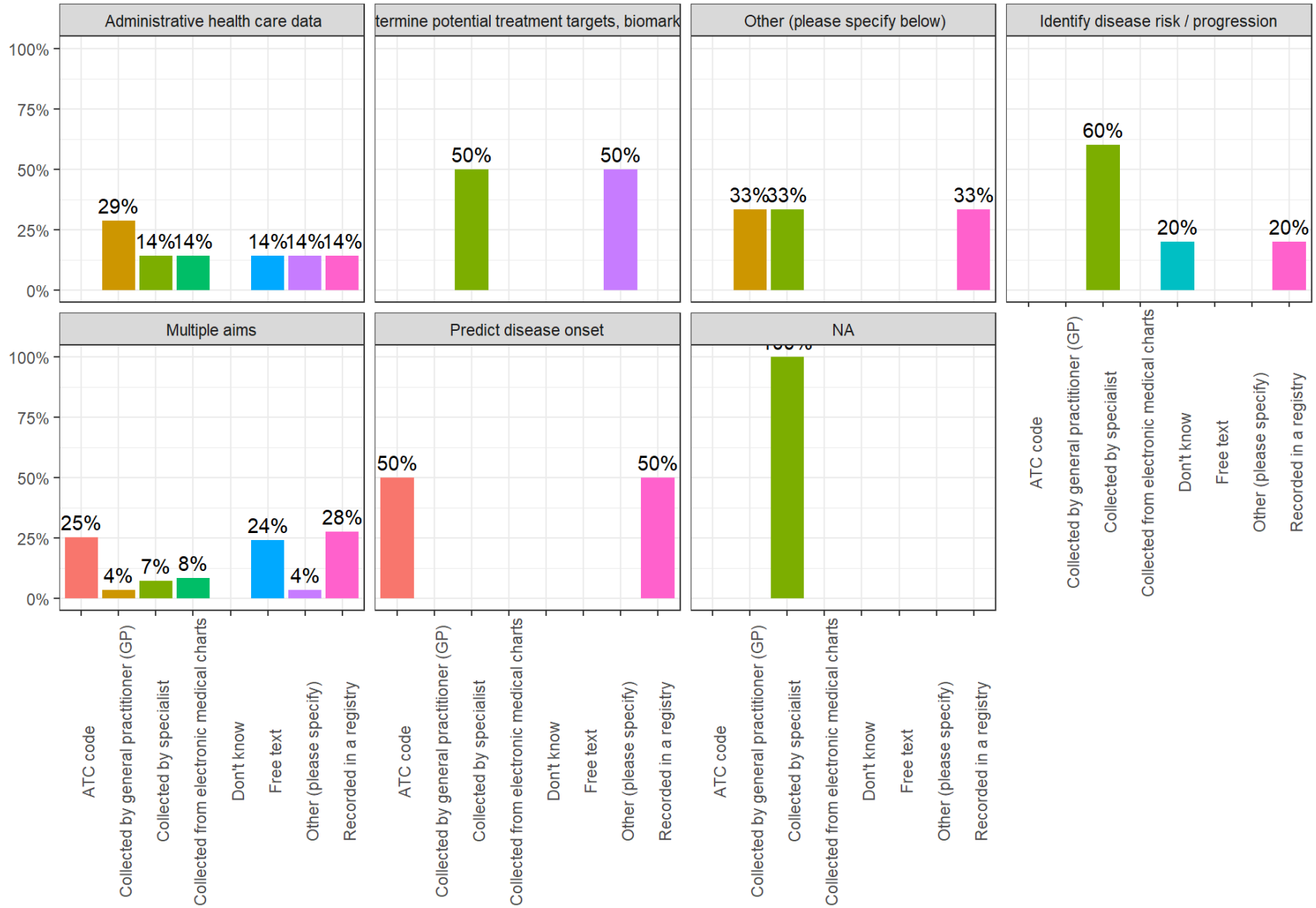
Answers divided by type of study:

| | Biobank (n=3) | Cohort study (n=25) | Consortium study (n=11) | Health administrative data (n=27) | Other (please specify below) (n=25) | Registry (n=12) | Overall (n=103) |
|---|---|---|---|---|---|---|---|
| **medications_data** | | | | | | | |
| ATC code | 1 (33.3%) | 3 (12.0%) | 1 (9.1%) | 9 (33.3%) | 8 (32.0%) | 0 (0%) | 22 (21.4%) |
| Other (please specify) | 1 (33.3%) | 1 (4.0%) | 1 (9.1%) | 0 (0%) | 0 (0%) | 2 (16.7%) | 5 (4.9%) |
| Recorded in a registry | 1 (33.3%) | 5 (20.0%) | 2 (18.2%) | 8 (29.6%) | 8 (32.0%) | 3 (25.0%) | 27 (26.2%) |
| Collected by general practitioner (GP) | 0 (0%) | 3 (12.0%) | 1 (9.1%) | 1 (3.7%) | 0 (0%) | 1 (8.3%) | 6 (5.8%) |
| Collected by specialist | 0 (0%) | 7 (28.0%) | 3 (27.3%) | 0 (0%) | 0 (0%) | 3 (25.0%) | 13 (12.6%) |
| Collected from electronic medical charts | 0 (0%) | 3 (12.0%) | 1 (9.1%) | 1 (3.7%) | 1 (4.0%) | 2 (16.7%) | 8 (7.8%) |
| Free text | 0 (0%) | 3 (12.0%) | 1 (9.1%) | 8 (29.6%) | 8 (32.0%) | 1 (8.3%) | 21 (20.4%) |
| Don't know | 0 (0%) | 0 (0%) | 1 (9.1%) | 0 (0%) | 0 (0%) | 0 (0%) | 1 (1.0%) |


Biobank
Cohort study
Consortium study
Health administrative data
Other (please specify below)
Registry
0%
10%
20%
30%
12%
28%
20%
9%
27%
18%
4%
30%
32%
8%
25%
17%
ATC code
Collected by general practitioner (GP)
Collected by specialist
Collected from electronic medical charts
Don't know
Free text
Other (please specify)
Recorded in a registry

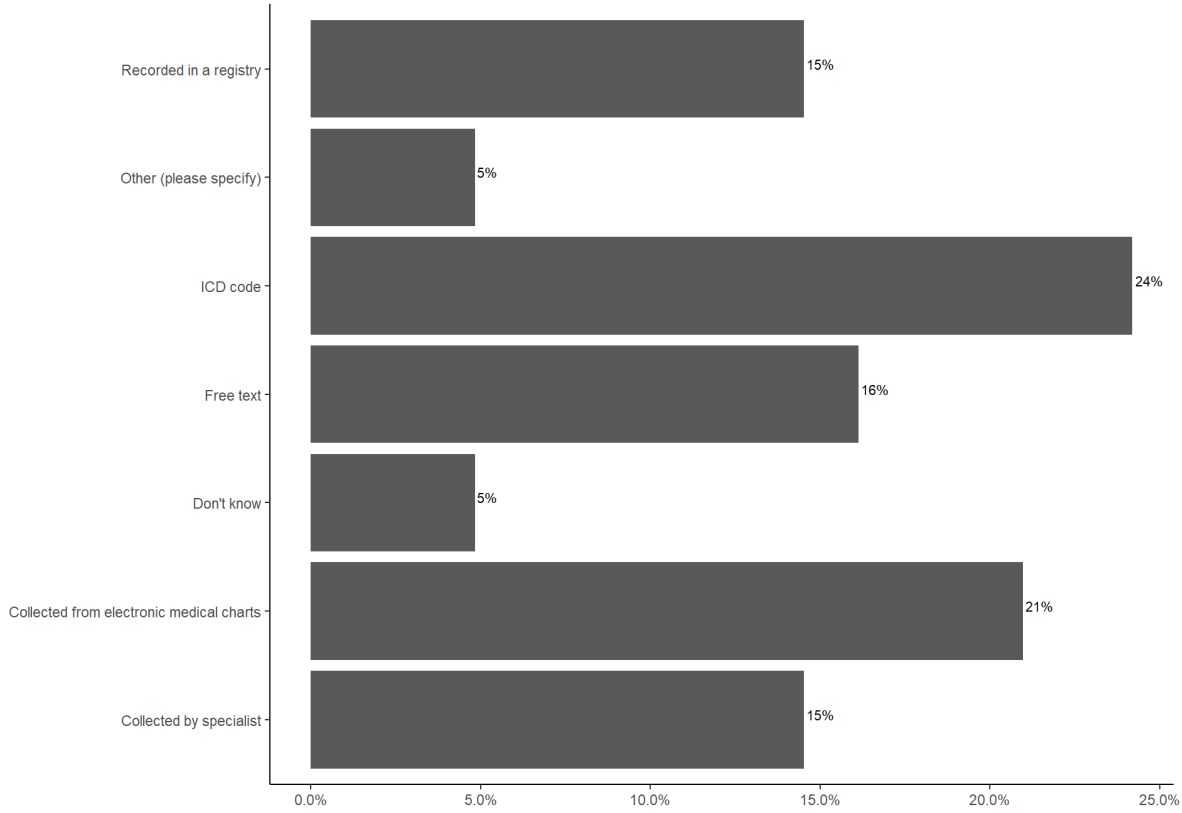
Answers divided by purpose of study:

| | Administrative health care data (n=7) | Determine potential treatment targets, biomarkers (n=2) | Other (please specify below) (n=3) | Identify disease risk / progression (n=5) | Multiple aims (n=83) | Predict disease onset (n=2) | Overall (n=103) |
|---|---|---|---|---|---|---|---|
| **medications_data** | | | | | | | |
| Collected by general practitioner (GP) | 2 (28.6%) | 0 (0%) | 1 (33.3%) | 0 (0%) | 3 (3.6%) | 0 (0%) | 6 (5.8%) |
| Collected by specialist | 1 (14.3%) | 1 (50.0%) | 1 (33.3%) | 3 (60.0%) | 6 (7.2%) | 0 (0%) | 13 (12.6%) |
| Collected from electronic medical charts | 1 (14.3%) | 0 (0%) | 0 (0%) | 0 (0%) | 7 (8.4%) | 0 (0%) | 8 (7.8%) |
| Free text | 1 (14.3%) | 0 (0%) | 0 (0%) | 0 (0%) | 20 (24.1%) | 0 (0%) | 21 (20.4%) |
| Other (please specify) | 1 (14.3%) | 1 (50.0%) | 0 (0%) | 0 (0%) | 3 (3.6%) | 0 (0%) | 5 (4.9%) |
| Recorded in a registry | 1 (14.3%) | 0 (0%) | 1 (33.3%) | 1 (20.0%) | 23 (27.7%) | 1 (50.0%) | 27 (26.2%) |
| Don't know | 0 (0%) | 0 (0%) | 0 (0%) | 1 (20.0%) | 0 (0%) | 0 (0%) | 1 (1.0%) |
| ATC code | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 21 (25.3%) | 1 (50.0%) | 22 (21.4%) |


Administrative health care data
termine potential treatment targets, biomark
Other (please specify below)
Identify disease risk / progression
Multiple aims
Predict disease onset
NA
100%
75%
50%
25%
0%
29%
14%14%
14%14%14%
50%
50%
33%33%
33%
60%
20%
20%
25%
4%
7%
8%
24%
4%
28%
50%
50%
ATC code
Collected by general practitioner (GP)
Collected by specialist
Collected from electronic medical charts
Don't know
Free text
Other (please specify)
Recorded in a registry

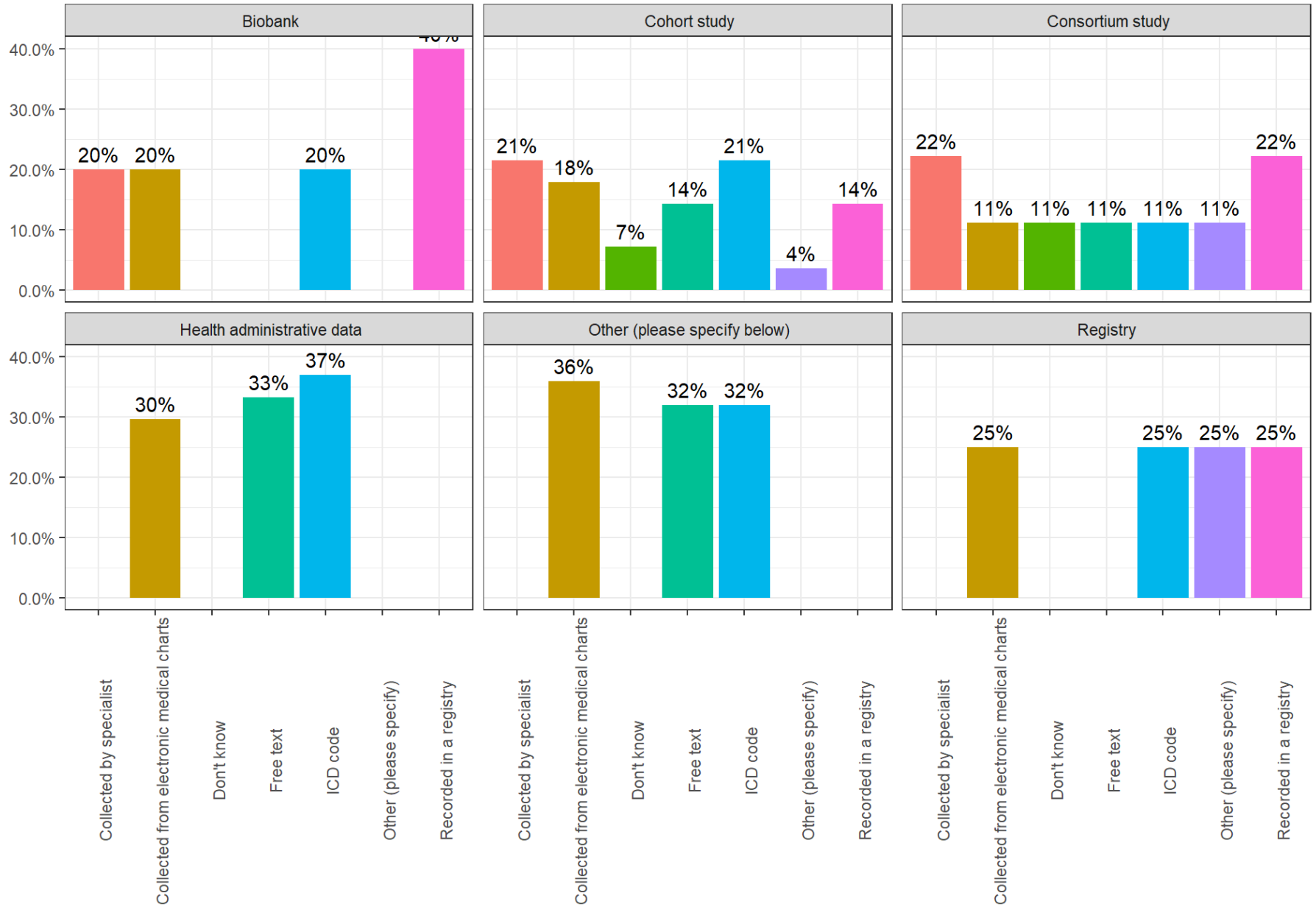
**20. How is the hospitalisation data being collected and from which source?**

| | Overall (n=62) |
|---|---|
| **How is the hopitalisation data being collected?** | |
| Collected by specialist | 9 (14.5%) |
| Collected from electronic medical charts | 13 (21.0%) |
| Don't know | 3 (4.8%) |
| Free text | 10 (16.1%) |
| ICD code | 15 (24.2%) |
| Other (please specify) | 3 (4.8%) |
| Recorded in a registry | 9 (14.5%) |



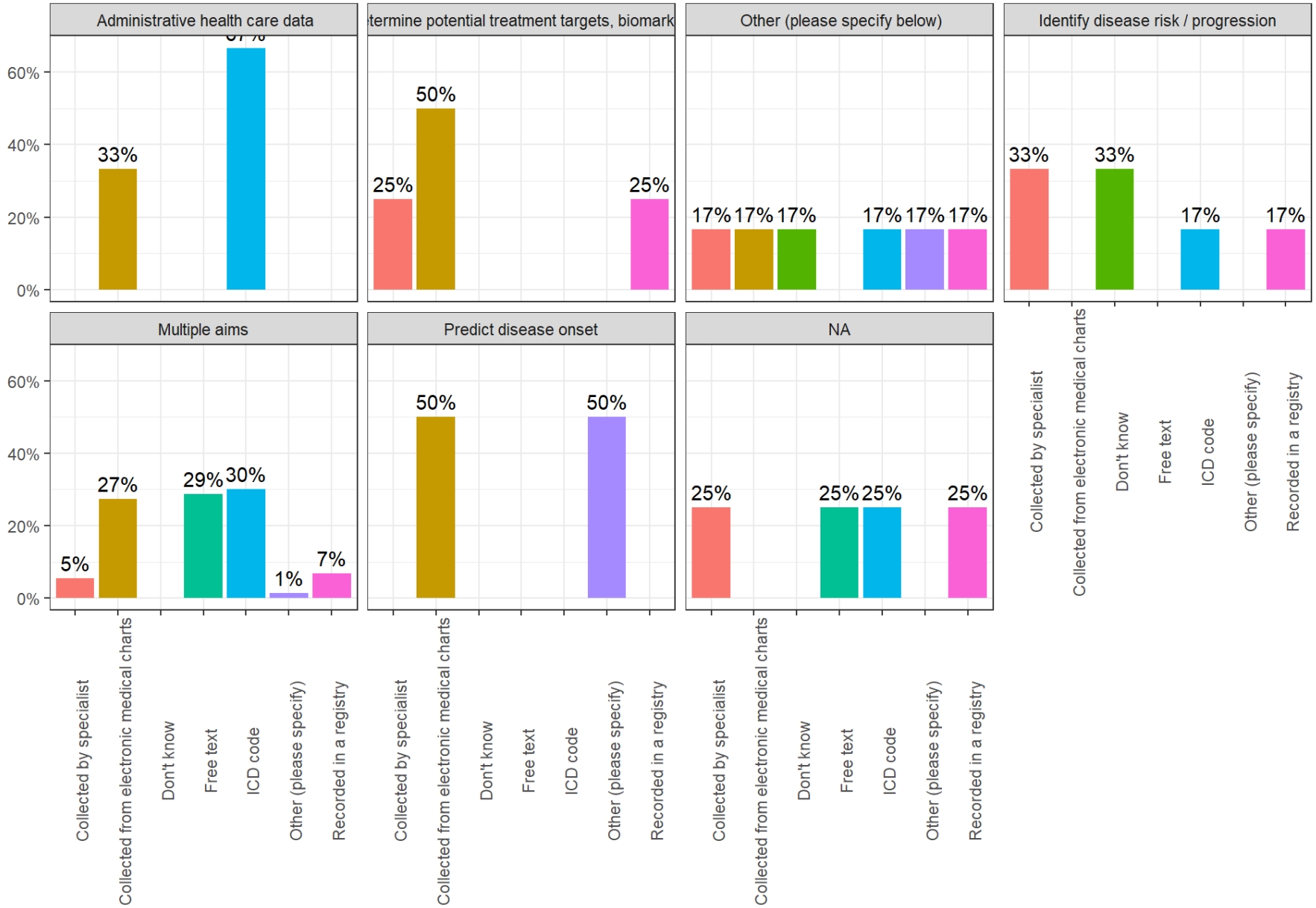
Answers divided by type of study:

| | Biobank (n=5) | Cohort study (n=28) | Consortium study (n=9) | Health administrative data (n=27) | Other (please specify below) (n=25) | Registry (n=4) | Overall (n=98) |
|---|---|---|---|---|---|---|---|
| **hospitalisation_data** | | | | | | | |
| Collected by specialist | 1 (20.0%) | 6 (21.4%) | 2 (22.2%) | 0 (0%) | 0 (0%) | 0 (0%) | 9 (9.2%) |
| Collected from electronic medical charts | 1 (20.0%) | 5 (17.9%) | 1 (11.1%) | 8 (29.6%) | 9 (36.0%) | 1 (25.0%) | 25 (25.5%) |
| ICD code | 1 (20.0%) | 6 (21.4%) | 1 (11.1%) | 10 (37.0%) | 8 (32.0%) | 1 (25.0%) | 27 (27.6%) |
| Recorded in a registry | 2 (40.0%) | 4 (14.3%) | 2 (22.2%) | 0 (0%) | 0 (0%) | 1 (25.0%) | 9 (9.2%) |
| Don't know | 0 (0%) | 2 (7.1%) | 1 (11.1%) | 0 (0%) | 0 (0%) | 0 (0%) | 3 (3.1%) |
| Free text | 0 (0%) | 4 (14.3%) | 1 (11.1%) | 9 (33.3%) | 8 (32.0%) | 0 (0%) | 22 (22.4%) |
| Other (please specify) | 0 (0%) | 1 (3.6%) | 1 (11.1%) | 0 (0%) | 0 (0%) | 1 (25.0%) | 3 (3.1%) |


Biobank
Cohort study
Consortium study
Health administrative data
Other (please specify below)
Registry
40.0%
30.0%
20.0%
10.0%
0.0%
20% 20% 20%
21% 18% 7% 14% 21% 4% 14%
22% 11% 11% 11% 11% 11% 22%
30% 33% 37%
36% 32% 32%
25% 25% 25% 25%
Collected by specialist
Collected from electronic medical charts
Don't know
Free text
ICD code
Other (please specify)
Recorded in a registry

Answers divided by purpose of study:

| | Administrative health care data (n=3) | Determine potential treatment targets, biomarkers (n=4) | Other (please specify below) (n=6) | Identify disease risk / progression (n=6) | Multiple aims (n=73) | Predict disease onset (n=2) | Overall (n=98) |
|---|---|---|---|---|---|---|---|
| **hospitalisation_data** | | | | | | | |
| Collected from electronic medical charts | 1 (33.3%) | 2 (50.0%) | 1 (16.7%) | 0 (0%) | 20 (27.4%) | 1 (50.0%) | 25 (25.5%) |
| ICD code | 2 (66.7%) | 0 (0%) | 1 (16.7%) | 1 (16.7%) | 22 (30.1%) | 0 (0%) | 27 (27.6%) |
| Collected by specialist | 0 (0%) | 1 (25.0%) | 1 (16.7%) | 2 (33.3%) | 4 (5.5%) | 0 (0%) | 9 (9.2%) |
| Recorded in a registry | 0 (0%) | 1 (25.0%) | 1 (16.7%) | 1 (16.7%) | 5 (6.8%) | 0 (0%) | 9 (9.2%) |
| Don't know | 0 (0%) | 0 (0%) | 1 (16.7%) | 2 (33.3%) | 0 (0%) | 0 (0%) | 3 (3.1%) |
| Other (please specify) | 0 (0%) | 0 (0%) | 1 (16.7%) | 0 (0%) | 1 (1.4%) | 1 (50.0%) | 3 (3.1%) |
| Free text | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 21 (28.8%) | 0 (0%) | 22 (22.4%) |


Administrative health care data
termine potential treatment targets, biomark
Other (please specify below)
Identify disease risk / progression
Multiple aims
Predict disease onset
NA
0%
20%
40%
60%
33%
25%
50%
25%
17%
17%
17%
17%
17%
17%
33%
33%
17%
17%
5%
27%
29%
30%
1%
7%
50%
50%
25%
25%
25%
25%
Collected by specialist
Collected from electronic medical charts
Don't know
Free text
ICD code
Other (please specify)
Recorded in a registry

**21. If data on hospitalisation is being recorded, is the reason for hospitalisation mentioned?**

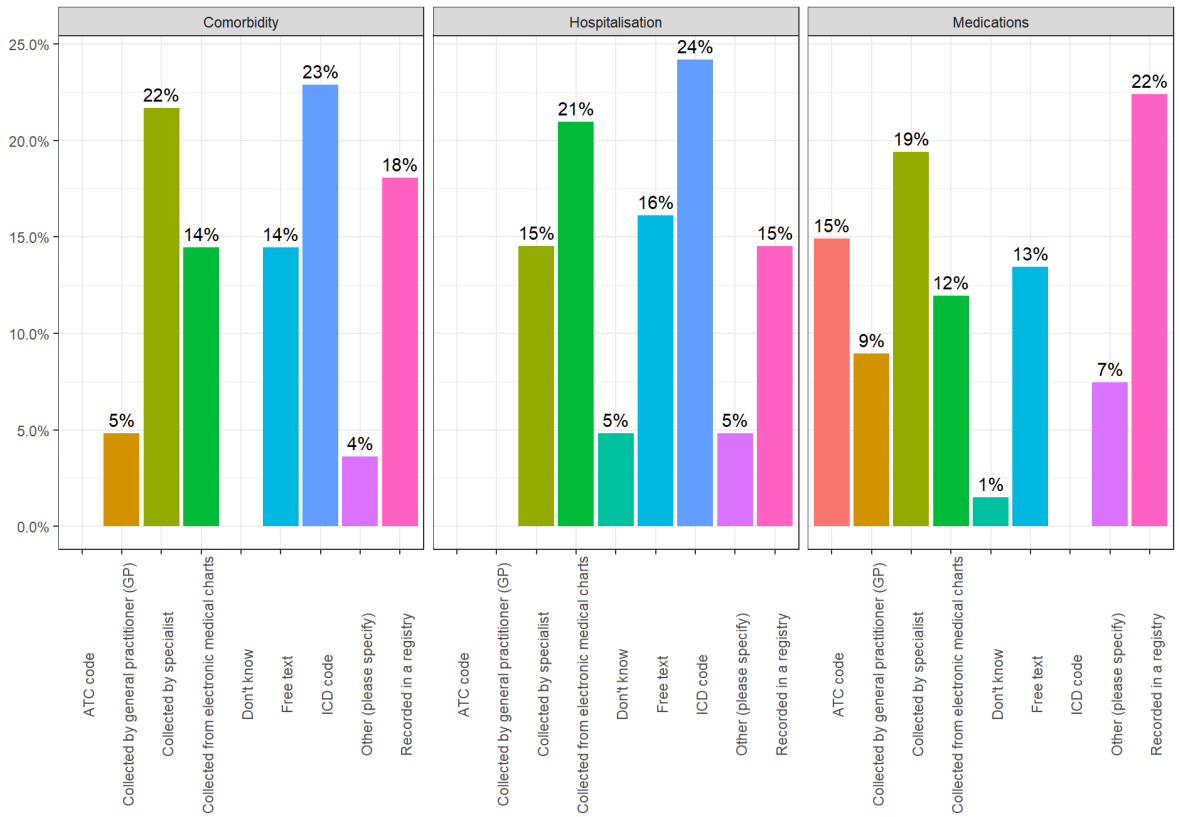
| | Overall (n=83) |
|---|---|
| **If data on hospitalisation is being recorded, is the reason for hospitalisation mentioned?** | |
| No | 11 (13.3%) |
| Yes | 19 (22.9%) |
| Missing | 53 (63.9%) |

**22. If data on hospitalisation is being recorded, is the length/date of stay mentioned?**

| | Overall (n=83) |
|---|---|
| **If data on hospitalisation is being recorded, is the length/date of stay mentioned?** | |
| No | 17 (20.5%) |
| Yes | 17 (20.5%) |
| Missing | 49 (59.0%) |

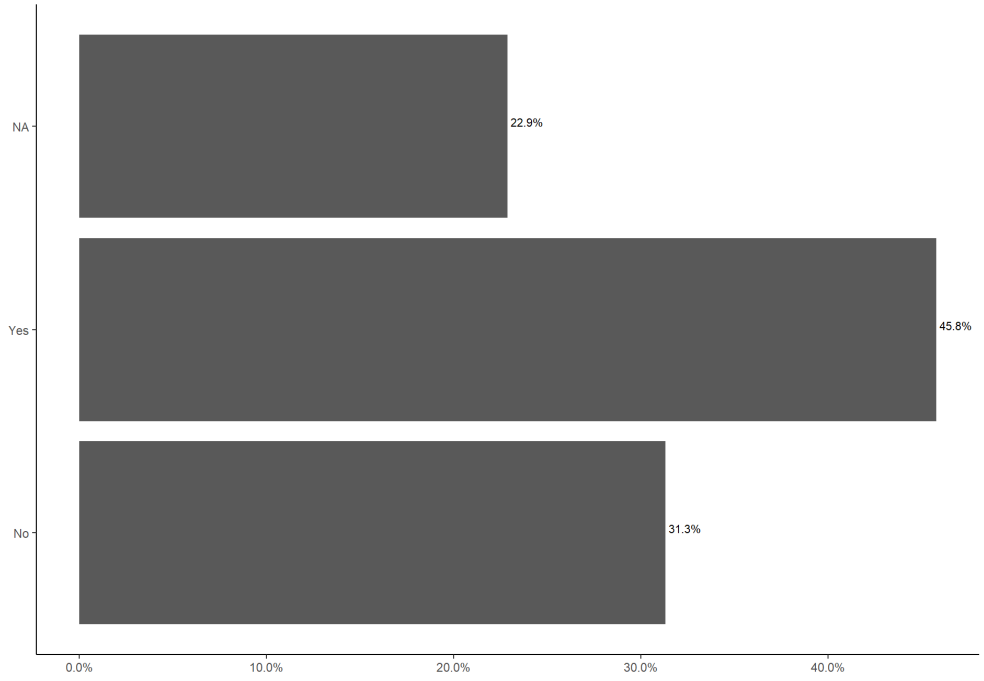
### 23. How is data being collected?

| | Comorbidity (n=83) | Hospitalisation (n=62) | Medications (n=67) | Overall (n=212) |
|---|---|---|---|---|
| Collected by general practitioner (GP) | 4 (4.8%) | 0 (0%) | 6 (9.0%) | 10 (4.7%) |
| Collected by specialist | 18 (21.7%) | 9 (14.5%) | 13 (19.4%) | 40 (18.9%) |
| Collected from electronic medical charts | 12 (14.5%) | 13 (21.0%) | 8 (11.9%) | 33 (15.6%) |
| Free text | 12 (14.5%) | 10 (16.1%) | 9 (13.4%) | 31 (14.6%) |
| ICD code | 19 (22.9%) | 15 (24.2%) | 0 (0%) | 34 (16.0%) |
| Other (please specify) | 3 (3.6%) | 3 (4.8%) | 5 (7.5%) | 11 (5.2%) |
| Recorded in a registry | 15 (18.1%) | 9 (14.5%) | 15 (22.4%) | 39 (18.4%) |
| Don't know | 0 (0%) | 3 (4.8%) | 1 (1.5%) | 4 (1.9%) |
| ATC code | 0 (0%) | 0 (0%) | 10 (14.9%) | 10 (4.7%) |



**24. Is disease specific clinical data being collected?**

| | Overall (n=83) |
|---|---|
| **Is disease specific clinical data being collected?** | |
| No | 26 (31.3%) |
| Yes | 38 (45.8%) |
| Missing | 19 (22.9%) |



Answer divided by type of study:

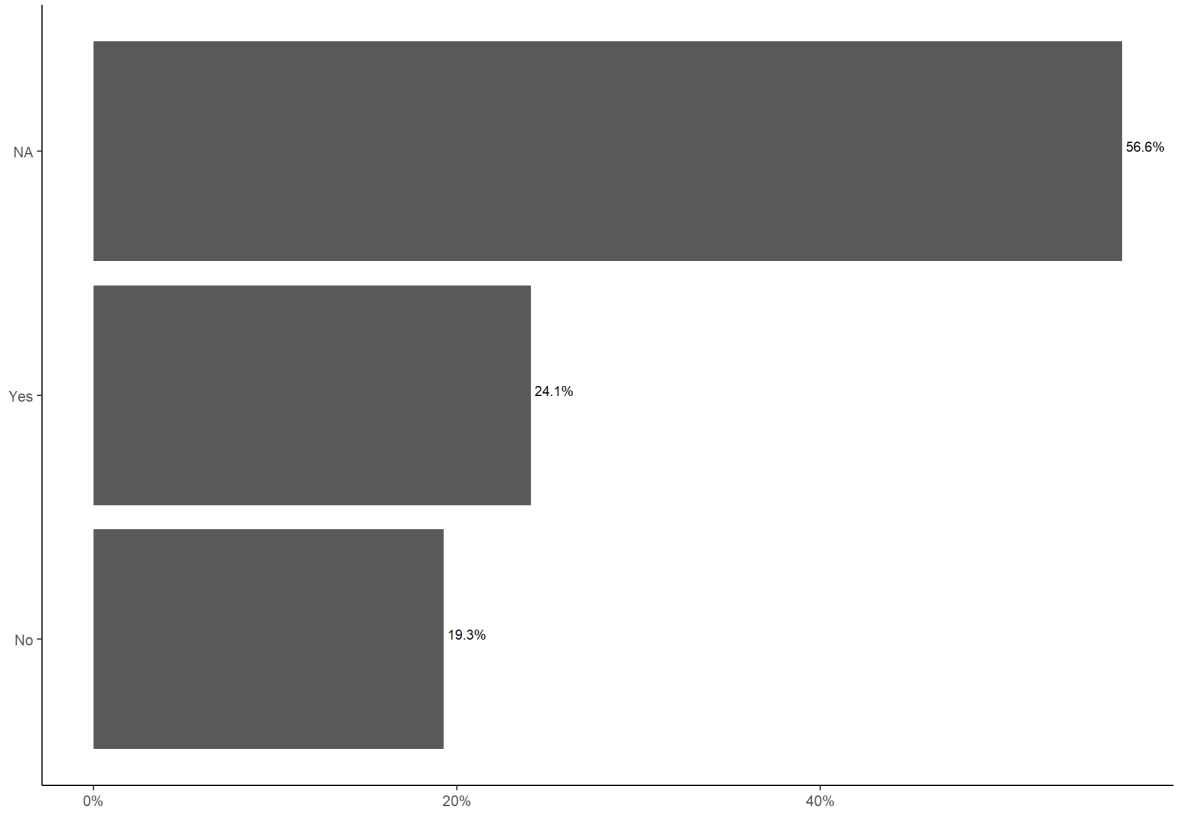
| | Biobank (n=4) | Cohort study (n=20) | Consortium study (n=15) | Health administrative data (n=4) | Other (please specify below) (n=12) | Registry (n=9) | Overall (n=83) |
|---|---|---|---|---|---|---|---|
| **Is disease specific clinical data being collected?** | | | | | | | |
| No | 1 (25.0%) | 4 (20.0%) | 5 (33.3%) | 3 (75.0%) | 7 (58.3%) | 4 (44.4%) | 26 (31.3%) |
| Yes | 2 (50.0%) | 15 (75.0%) | 10 (66.7%) | 1 (25.0%) | 4 (33.3%) | 5 (55.6%) | 38 (45.8%) |
| Missing | 1 (25.0%) | 1 (5.0%) | 0 (0%) | 0 (0%) | 1 (8.3%) | 0 (0%) | 19 (22.9%) |

Answer divided by purpose of study:

| | Administrative health care data (n=4) | Determine potential treatment targets, biomarkers (n=4) | Don't know (n=1) | Other (please specify below) (n=3) | Patient/cohort classification (n=1) | Determine effects of treatment (n=1) | Identify disease risk / progression (n=6) | Multiple aims (n=37) | Predict disease onset (n=4) | Overall (n=83) |
|---|---|---|---|---|---|---|---|---|---|---|
| **Is disease specific clinical data being collected?** | | | | | | | | | | |
| No | 1 (25.0%) | 1 (25.0%) | 1 (100%) | 1 (33.3%) | 1 (100%) | 1 (100%) | 2 (33.3%) | 15 (40.5%) | 1 (25.0%) | 26 (31.3%) |
| Yes | 3 (75.0%) | 3 (75.0%) | 0 (0%) | 2 (66.7%) | 0 (0%) | 0 (0%) | 4 (66.7%) | 22 (59.5%) | 2 (50.0%) | 38 (45.8%) |
| Missing | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 1 (25.0%) | 19 (22.9%) |

**25. Is the date of disease onset (first disease manifestation or symptoms) being reported?**

| | Overall (n=83) |
|---|---|
| **Is the date of disease onset (first disease manifestation or symptoms) being reported?** | |
| No | 16 (19.3%) |
| Yes | 20 (24.1%) |
| Missing | 47 (56.6%) |



Answer divided by type of study:

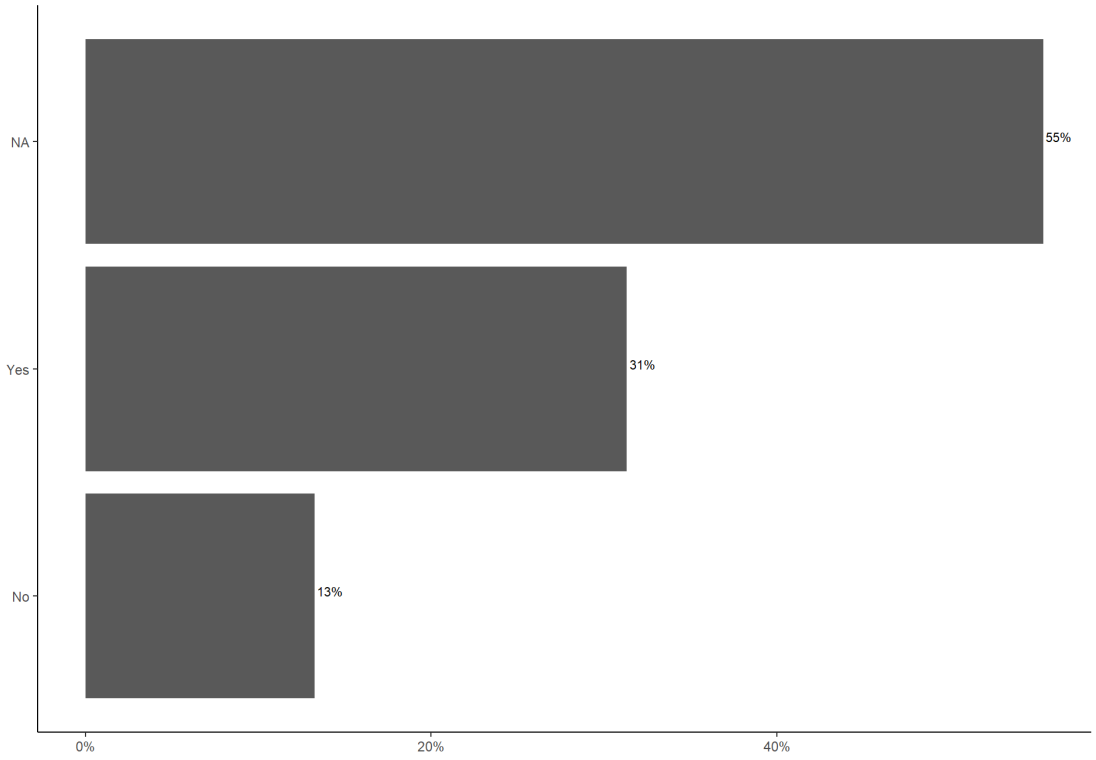
| | Biobank (n=4) | Cohort study (n=20) | Consortium study (n=15) | Health administrative data (n=4) | Other (please specify below) (n=12) | Registry (n=9) | Overall (n=83) |
|---|---|---|---|---|---|---|---|
| **Is the date of disease onset (first disease manifestation or symptoms) being reported?** | | | | | | | |
| No | 1 (25.0%) | 6 (30.0%) | 3 (20.0%) | 1 (25.0%) | 3 (25.0%) | 1 (11.1%) | 16 (19.3%) |
| Yes | 1 (25.0%) | 8 (40.0%) | 6 (40.0%) | 0 (0%) | 1 (8.3%) | 4 (44.4%) | 20 (24.1%) |
| Missing | 2 (50.0%) | 6 (30.0%) | 6 (40.0%) | 3 (75.0%) | 8 (66.7%) | 4 (44.4%) | 47 (56.6%) |

Answer divided by purpose of study:

| | Administrative health care data (n=4) | Determine potential treatment targets, biomarkers (n=4) | Don't know (n=1) | Other (please specify below) (n=3) | Patient/cohort classification (n=1) | Determine effects of treatment (n=1) | Identify disease risk / progression (n=6) | Multiple aims (n=37) | Predict disease onset (n=4) | Overall (n=83) |
|---|---|---|---|---|---|---|---|---|---|---|
| **Is the date of disease onset (first disease manifestation or symptoms) being reported?** | | | | | | | | | | |
| No | 2 (50.0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 1 (16.7%) | 12 (32.4%) | 0 (0%) | 16 (19.3%) |
| Yes | 1 (25.0%) | 2 (50.0%) | 0 (0%) | 2 (66.7%) | 0 (0%) | 0 (0%) | 3 (50.0%) | 9 (24.3%) | 2 (50.0%) | 20 (24.1%) |
| Missing | 1 (25.0%) | 2 (50.0%) | 1 (100%) | 1 (33.3%) | 1 (100%) | 1 (100%) | 2 (33.3%) | 16 (43.2%) | 2 (50.0%) | 47 (56.6%) |

**26. Is the date of diagnosis being reported?**

| | Overall (n=83) |
|---|---|
| **Is the date of diagnosis being reported?** | |
| No | 11 (13.3%) |
| Yes | 26 (31.3%) |
| Missing | 46 (55.4%) |



Answers divided by type of study:

| | Biobank (n=4) | Cohort study (n=20) | Consortium study (n=15) | Health administrative data (n=4) | Other (please specify below) (n=12) | Registry (n=9) | Overall (n=83) |
|---|---|---|---|---|---|---|---|
| **Is the date of diagnosis being reported?** | | | | | | | |
| Yes | 2 (50.0%) | 10 (50.0%) | 6 (40.0%) | 1 (25.0%) | 3 (25.0%) | 4 (44.4%) | 26 (31.3%) |
| No | 0 (0%) | 5 (25.0%) | 3 (20.0%) | 0 (0%) | 1 (8.3%) | 1 (11.1%) | 11 (13.3%) |
| Missing | 2 (50.0%) | 5 (25.0%) | 6 (40.0%) | 3 (75.0%) | 8 (66.7%) | 4 (44.4%) | 46 (55.4%) |

Answers divided by purpose of study:

| | Administrative health care data (n=4) | Determine potential treatment targets, biomarkers (n=4) | Don't know (n=1) | Other (please specify below) (n=3) | Patient/cohort classification (n=1) | Determine effects of treatment (n=1) | Identify disease risk / progression (n=6) | Multiple aims (n=37) | Predict disease onset (n=4) | Overall (n=83) |
|---|---|---|---|---|---|---|---|---|---|---|
| **Is the date of diagnosis being reported?** | | | | | | | | | | |
| No | 1 (25.0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 1 (16.7%) | 8 (21.6%) | 1 (25.0%) | 11 (13.3%) |
| Yes | 2 (50.0%) | 2 (50.0%) | 0 (0%) | 2 (66.7%) | 0 (0%) | 0 (0%) | 3 (50.0%) | 14 (37.8%) | 1 (25.0%) | 26 (31.3%) |
| Missing | 1 (25.0%) | 2 (50.0%) | 1 (100%) | 1 (33.3%) | 1 (100%) | 1 (100%) | 2 (33.3%) | 15 (40.5%) | 2 (50.0%) | 46 (55.4%) |

**27. Are the specific disease's severity measures being reported?**

| | Overall (n=83) |
|---|---|
| **Are the specific disease's severity measures being reported?** | |
| No | 8 (9.6%) |
| Yes | 28 (33.7%) |
| Missing | 47 (56.6%) |

Answers divided by type of study:

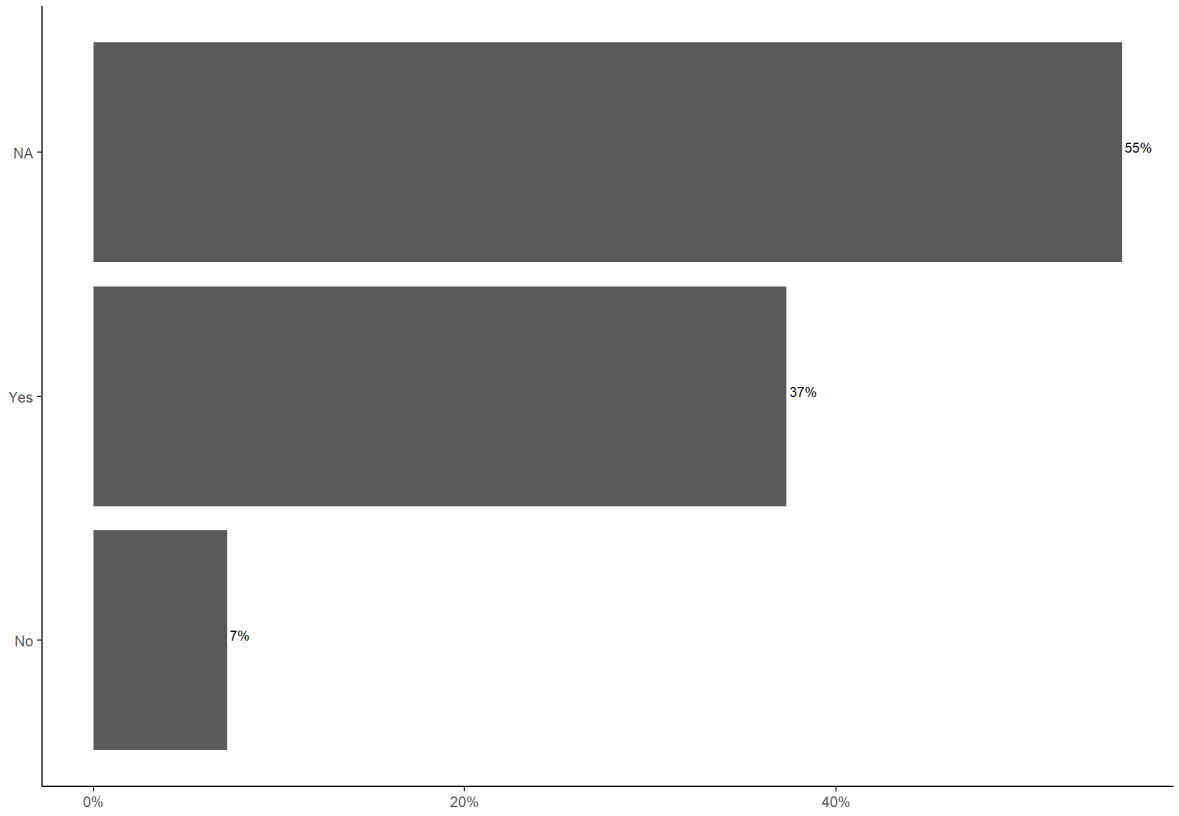
| | Biobank (n=4) | Cohort study (n=20) | Consortium study (n=15) | Health administrative data (n=4) | Other (please specify below) (n=12) | Registry (n=9) | Overall (n=83) |
|---|---|---|---|---|---|---|---|
| **Are the specific disease's severity measures being reported?** | | | | | | | |
| Yes | 2 (50.0%) | 13 (65.0%) | 6 (40.0%) | 0 (0%) | 3 (25.0%) | 3 (33.3%) | 28 (33.7%) |
| No | 0 (0%) | 2 (10.0%) | 2 (13.3%) | 1 (25.0%) | 1 (8.3%) | 2 (22.2%) | 8 (9.6%) |
| Missing | 2 (50.0%) | 5 (25.0%) | 7 (46.7%) | 3 (75.0%) | 8 (66.7%) | 4 (44.4%) | 47 (56.6%) |

Answers divided by purpose of study:

| | Administrative health care data (n=4) | Determine potential treatment targets, biomarkers (n=4) | Don't know (n=1) | Other (please specify below) (n=3) | Patient/cohort classification (n=1) | Determine effects of treatment (n=1) | Identify disease risk / progression (n=6) | Multiple aims (n=37) | Predict disease onset (n=4) | Overall (n=83) |
|---|---|---|---|---|---|---|---|---|---|---|
| **Are the specific disease's severity measures being reported?** | | | | | | | | | | |
| No | 3 (75.0%) | 0 (0%) | 0 (0%) | 1 (33.3%) | 0 (0%) | 0 (0%) | 1 (16.7%) | 3 (8.1%) | 0 (0%) | 8 (9.6%) |
| Yes | 0 (0%) | 2 (50.0%) | 0 (0%) | 1 (33.3%) | 0 (0%) | 0 (0%) | 3 (50.0%) | 18 (48.6%) | 2 (50.0%) | 28 (33.7%) |
| Missing | 1 (25.0%) | 2 (50.0%) | 1 (100%) | 1 (33.3%) | 1 (100%) | 1 (100%) | 2 (33.3%) | 16 (43.2%) | 2 (50.0%) | 47 (56.6%) |

### 28. Are the specific disease's treatments being reported?

| | Overall (n=83) |
|---|---|
| **Are the specific disease's treatments being reported?** | |
| No | 6 (7.2%) |
| Yes | 31 (37.3%) |
| Missing | 46 (55.4%) |



## Which treatment is reported?

| | Overall (n=83) |
|---|---|
| **VAR50C** | |
| ?? via normal procedure codes | 1 (1.2%) |
| All pharmacologic treatment. Focus on bDMARD and tsDMARD | 1 (1.2%) |
| Clinical trial. All interventions, therapeutic and retaliative are recorded. | 1 (1.2%) |
| Disease modifying treatment is captured | 1 (1.2%) |
| diverse, depends on cancer type. | 1 (1.2%) |
| e.g., antibiotics | 1 (1.2%) |
| From registries | 1 (1.2%) |
| Gastric bypass surgery, medication to treat comorbidities (diabetes, dyslipidemia, hypertension, etc) | 1 (1.2%) |
| immunosuppressive therapy | 1 (1.2%) |
| In some cohorts, the treatments given to Inflammatory Bowel Disease patients are reported. | 1 (1.2%) |
| In structured text and free text. | 1 (1.2%) |
| interferon beta, glatiramer acetate, dimethyl-fumarate, teriflunomide, fingolimod, natalizumab, ocrelizumab, rituximab, alemtuzumab, cladribine | 1 (1.2%) |
| medication | 1 (1.2%) |
| medication, non-invasive brain stimulation | 1 (1.2%) |
| Naturalistic, pharmacological interventions are systematically registered | 1 (1.2%) |
| Only primary care prescription data is being captured, hospital issued prescriptions are not available. | 1 (1.2%) |
| Potentially through health records | 1 (1.2%) |
| Procedures are reported in PEDW. We have GP prescribing. | 1 (1.2%) |
| sometimes info on RT, CTX | 4 (4.8%) |
| Surgery and/or RT/CRT | 1 (1.2%) |
| Treatments for Parkinson's disease | 1 (1.2%) |
| Missing | 59 (71.1%) |

Answers divided by type of study:

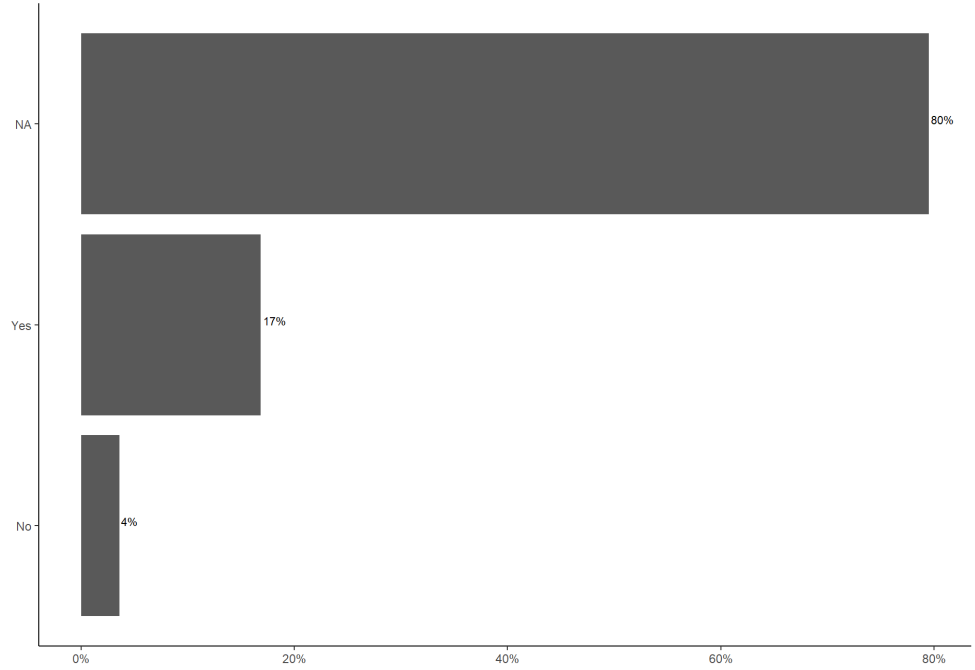
| | Biobank (n=4) | Cohort study (n=20) | Consortium study (n=15) | Health administrative data (n=4) | Other (please specify below) (n=12) | Registry (n=9) | Overall (n=83) |
|---|---|---|---|---|---|---|---|
| **Are the specific disease's treatments being reported?** | | | | | | | |
| Yes | 2 (50.0%) | 14 (70.0%) | 7 (46.7%) | 1 (25.0%) | 4 (33.3%) | 3 (33.3%) | 31 (37.3%) |
| No | 0 (0%) | 1 (5.0%) | 2 (13.3%) | 0 (0%) | 0 (0%) | 2 (22.2%) | 6 (7.2%) |
| Missing | 2 (50.0%) | 5 (25.0%) | 6 (40.0%) | 3 (75.0%) | 8 (66.7%) | 4 (44.4%) | 46 (55.4%) |

Answers divided by purpose of study:

| | Administrative health care data (n=4) | Determine potential treatment targets, biomarkers (n=4) | Don't know (n=1) | Other (please specify below) (n=3) | Patient/cohort classification (n=1) | Determine effects of treatment (n=1) | Identify disease risk / progression (n=6) | Multiple aims (n=37) | Predict disease onset (n=4) | Overall (n=83) |
|---|---|---|---|---|---|---|---|---|---|---|
| **Are the specific disease's treatments being reported?** | | | | | | | | | | |
| No | 2 (50.0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 3 (8.1%) | 1 (25.0%) | 6 (7.2%) |
| Yes | 1 (25.0%) | 2 (50.0%) | 0 (0%) | 2 (66.7%) | 0 (0%) | 0 (0%) | 4 (66.7%) | 19 (51.4%) | 1 (25.0%) | 31 (37.3%) |
| Missing | 1 (25.0%) | 2 (50.0%) | 1 (100%) | 1 (33.3%) | 1 (100%) | 1 (100%) | 2 (33.3%) | 15 (40.5%) | 2 (50.0%) | 46 (55.4%) |

## 29. Is treatment response (in any form) being reported?

| | Overall (n=83) |
|---|---|
| **Is treatment response (in any form) being reported?** | |
| No | 3 (3.6%) |
| Yes | 14 (16.9%) |
| Missing | 66 (79.5%) |



What is being reported?

| | Overall (n=83) |
|---|---|
| **VAR51C** | |
| Clinical scales measuring the reduction in severity of depressive epidode after 5 weeks | 1 (1.2%) |
| Disease activity measurements | 1 (1.2%) |
| for some substudies, overall survival or PFS are given (not always accurate, though) | 4 (4.8%) |
| Induction of remission after 6-14 weeks after treatment start is reported. | 1 (1.2%) |
| partially | 1 (1.2%) |
| radiological, ultrasound, Ki67 | 1 (1.2%) |
| scales | 1 (1.2%) |
| survival | 1 (1.2%) |
| Treatment response in form of MRI and disease severity | 1 (1.2%) |
| Missing | 71 (85.5%) |

Answers divided by type of study:

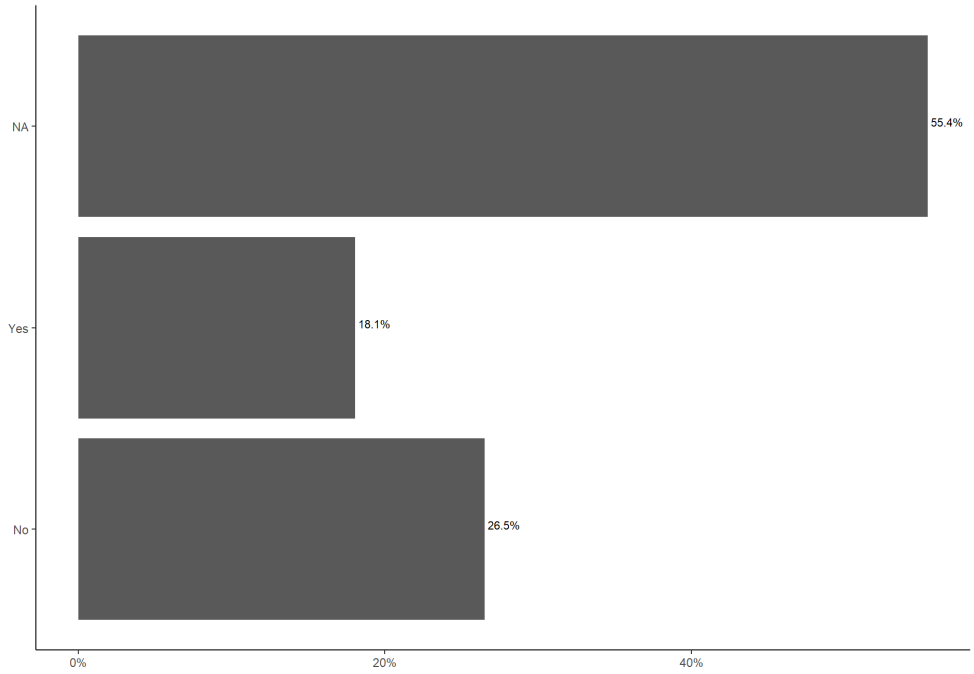
| | Biobank (n=4) | Cohort study (n=20) | Consortium study (n=15) | Health administrative data (n=4) | Other (please specify below) (n=12) | Registry (n=9) | Overall (n=83) |
|---|---|---|---|---|---|---|---|
| **Is treatment response (in any form) being reported?** | | | | | | | |
| Yes | 1 (25.0%) | 4 (20.0%) | 5 (33.3%) | 0 (0%) | 3 (25.0%) | 1 (11.1%) | 14 (16.9%) |
| No | 0 (0%) | 1 (5.0%) | 1 (6.7%) | 0 (0%) | 0 (0%) | 0 (0%) | 3 (3.6%) |
| Missing | 3 (75.0%) | 15 (75.0%) | 9 (60.0%) | 4 (100%) | 9 (75.0%) | 8 (88.9%) | 66 (79.5%) |

Answers divided by purpose of study:

| | Administrative health care data (n=4) | Determine potential treatment targets, biomarkers (n=4) | Don't know (n=1) | Other (please specify below) (n=3) | Patient/cohort classification (n=1) | Determine effects of treatment (n=1) | Identify disease risk / progression (n=6) | Multiple aims (n=37) | Predict disease onset (n=4) | Overall (n=83) |
|---|---|---|---|---|---|---|---|---|---|---|
| **Is treatment response (in any form) being reported?** | | | | | | | | | | |
| Yes | 0 (0%) | 1 (25.0%) | 0 (0%) | 1 (33.3%) | 0 (0%) | 0 (0%) | 1 (16.7%) | 10 (27.0%) | 1 (25.0%) | 14 (16.9%) |
| No | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 3 (8.1%) | 0 (0%) | 3 (3.6%) |
| Missing | 4 (100%) | 3 (75.0%) | 1 (100%) | 2 (66.7%) | 1 (100%) | 1 (100%) | 5 (83.3%) | 24 (64.9%) | 3 (75.0%) | 66 (79.5%) |

## 30. Are adverse events being recorded?

| | Overall (n=83) |
|---|---|
| **Are adverse events being recorded?** | |
| No | 22 (26.5%) |
| Yes | 15 (18.1%) |
| Missing | 46 (55.4%) |




Comments to answers:

| | Overall (n=83) |
|---|---|
| **VAR52C** | |
| Also during follow-up visits | 1 (1.2%) |
| as text | 1 (1.2%) |
| Clinical trial. In order to correlate adverse events with dosing or active principle. | 1 (1.2%) |
| free text | 1 (1.2%) |
| From registries | 1 (1.2%) |
| Only eventually reported | 1 (1.2%) |
| Only if recorded in the primary and secondary care record and coded as such. | 1 (1.2%) |
| patient diary | 1 (1.2%) |
| Potentially if reported in health records | 1 (1.2%) |
| They are not reported explicitly but may be extracted from the free text in EHRs. | 1 (1.2%) |
| Toxicity | 1 (1.2%) |
| Missing | 72 (86.7%) |

Answers divided by type of study:

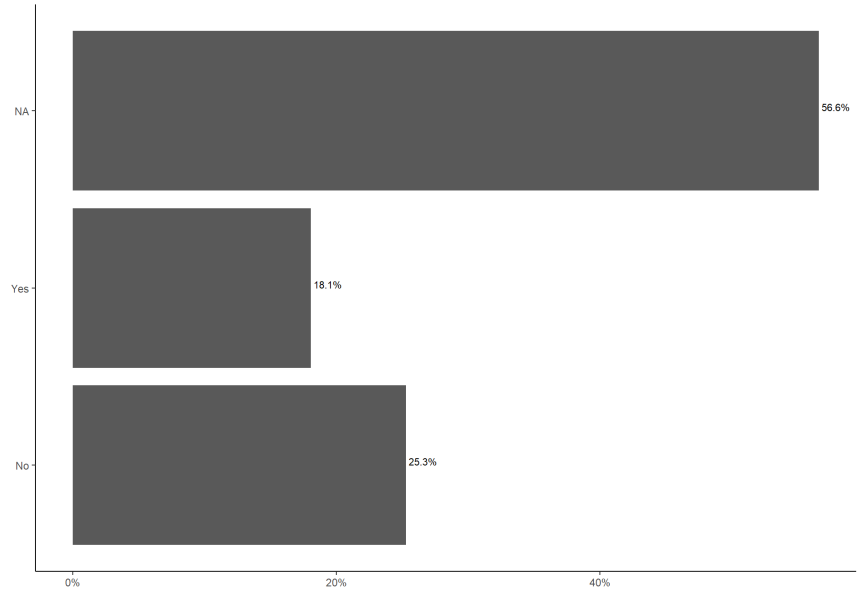
| | Biobank (n=4) | Cohort study (n=20) | Consortium study (n=15) | Health administrative data (n=4) | Other (please specify below) (n=12) | Registry (n=9) | Overall (n=83) |
|---|---|---|---|---|---|---|---|
| **Are adverse events being recorded?** | | | | | | | |
| No | 1 (25.0%) | 7 (35.0%) | 6 (40.0%) | 1 (25.0%) | 3 (25.0%) | 3 (33.3%) | 22 (26.5%) |
| Yes | 1 (25.0%) | 8 (40.0%) | 3 (20.0%) | 0 (0%) | 1 (8.3%) | 2 (22.2%) | 15 (18.1%) |
| Missing | 2 (50.0%) | 5 (25.0%) | 6 (40.0%) | 3 (75.0%) | 8 (66.7%) | 4 (44.4%) | 46 (55.4%) |

Answers divided by purpose of study:

| | Administrative health care data (n=4) | Determine potential treatment targets, biomarkers (n=4) | Don't know (n=1) | Other (please specify below) (n=3) | Patient/cohort classification (n=1) | Determine effects of treatment (n=1) | Identify disease risk / progression (n=6) | Multiple aims (n=37) | Predict disease onset (n=4) | Overall (n=83) |
|---|---|---|---|---|---|---|---|---|---|---|
| **Are adverse events being recorded?** | | | | | | | | | | |
| No | 3 (75.0%) | 2 (50.0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 2 (33.3%) | 12 (32.4%) | 2 (50.0%) | 22 (26.5%) |
| Yes | 0 (0%) | 0 (0%) | 0 (0%) | 2 (66.7%) | 0 (0%) | 0 (0%) | 2 (33.3%) | 10 (27.0%) | 0 (0%) | 15 (18.1%) |
| Missing | 1 (25.0%) | 2 (50.0%) | 1 (100%) | 1 (33.3%) | 1 (100%) | 1 (100%) | 2 (33.3%) | 15 (40.5%) | 2 (50.0%) | 46 (55.4%) |

### 31. Is paraclinical data being collected?

| | Overall (n=83) |
|---|---|
| **Is paraclinical data being collected?** | |
| No | 21 (25.3%) |
| Yes | 15 (18.1%) |
| Missing | 47 (56.6%) |




What type of paraclinical data is being collected?

| | Overall (n=83) |
|---|---|
| **VAR53C** | |
| All kinds of data related to hospital admission | 1 (1.2%) |
| Bioanalytical, radiological etc | 1 (1.2%) |
| Blood, serum and urine chemistry, immunological data. | 1 (1.2%) |
| brain MRI and retina OCT | 1 (1.2%) |
| Country, Hospital of treatment | 1 (1.2%) |
| CSF measures: OCB pos/neg, IgG index | 1 (1.2%) |
| f/MRI | 2 (2.4%) |
| From registries | 1 (1.2%) |
| MRI, biomarker levels (eg. NfL) | 1 (1.2%) |
| Quality of life (questionnaires) | 1 (1.2%) |
| We receive hospital lab reports and will have the all Wales data when available | 1 (1.2%) |
| Missing | 71 (85.5%) |

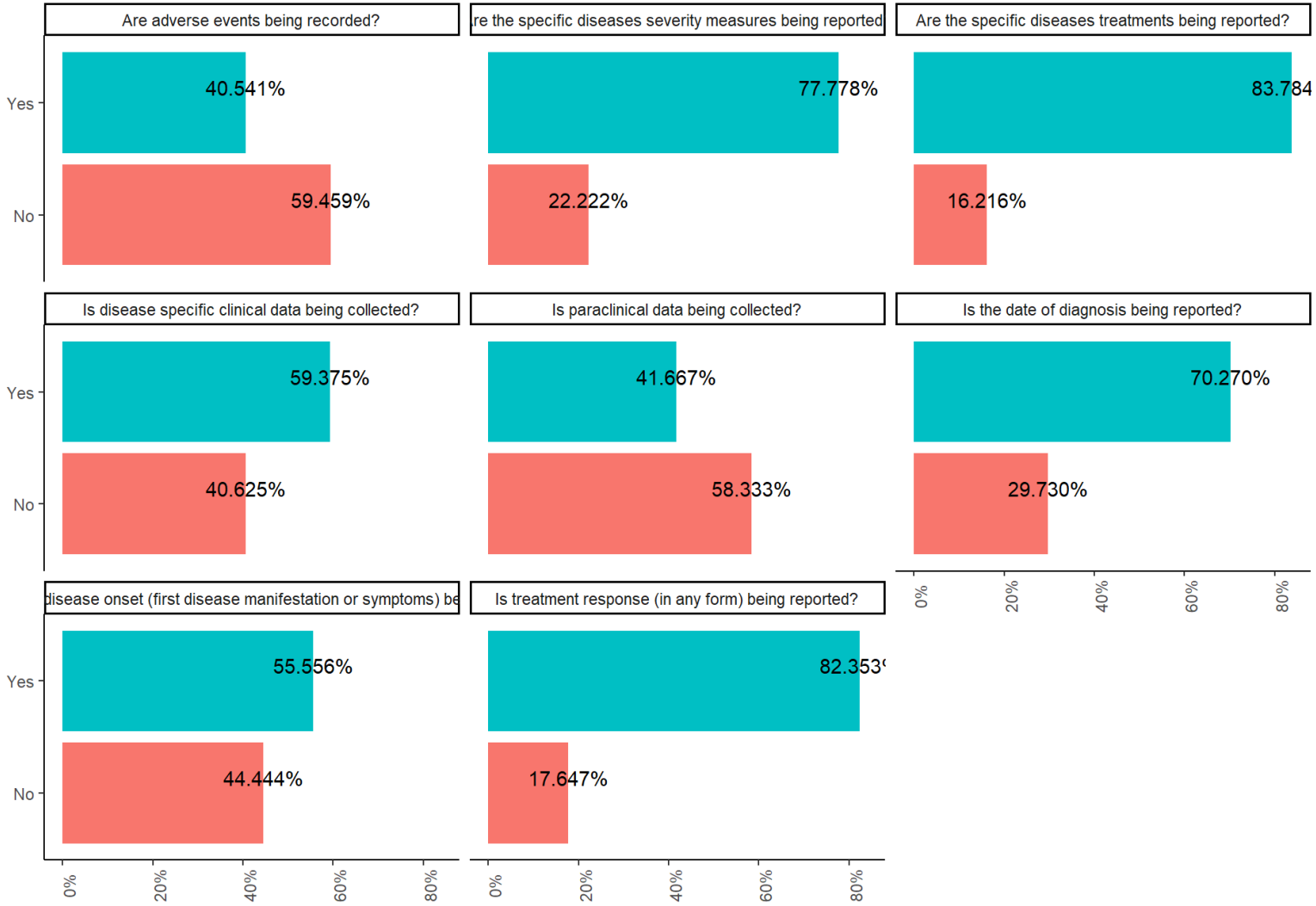
Answers divided by type of study:

| | Biobank (n=4) | Cohort study (n=20) | Consortium study (n=15) | Health administrative data (n=4) | Other (please specify below) (n=12) | Registry (n=9) | Overall (n=83) |
|---|---|---|---|---|---|---|---|
| **Is paraclinical data being collected?** | | | | | | | |
| No | 1 (25.0%) | 8 (40.0%) | 5 (33.3%) | 1 (25.0%) | 3 (25.0%) | 3 (33.3%) | 21 (25.3%) |
| Yes | 1 (25.0%) | 6 (30.0%) | 4 (26.7%) | 0 (0%) | 1 (8.3%) | 2 (22.2%) | 15 (18.1%) |
| Missing | 2 (50.0%) | 6 (30.0%) | 6 (40.0%) | 3 (75.0%) | 8 (66.7%) | 4 (44.4%) | 47 (56.6%) |

Answers divided by purpose of study:

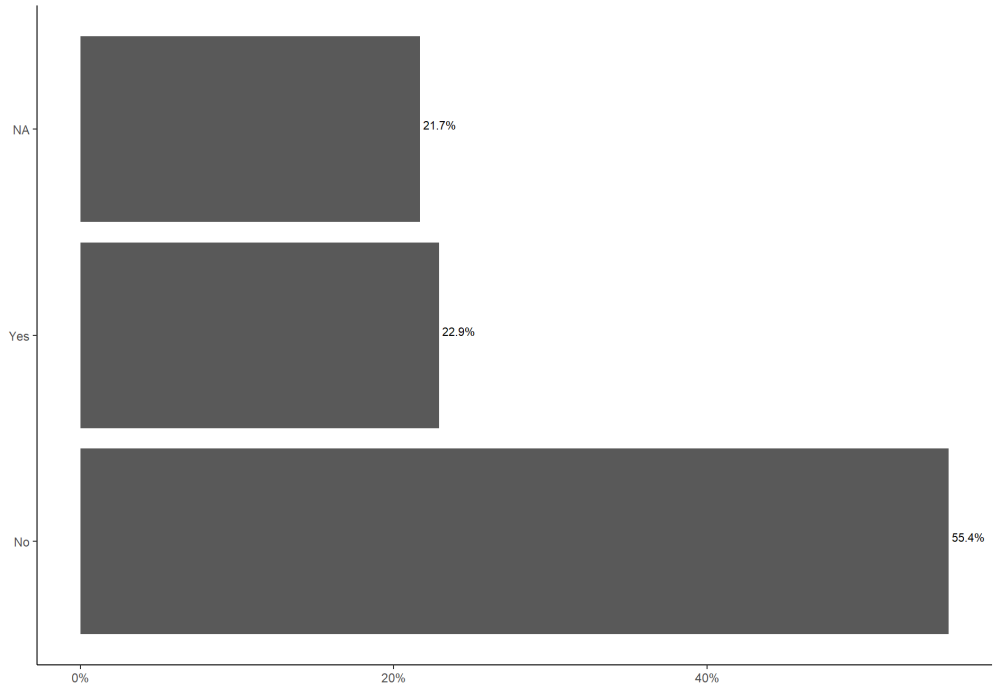
| | Administrative health care data (n=4) | Determine potential treatment targets, biomarkers (n=4) | Don't know (n=1) | Other (please specify below) (n=3) | Patient/cohort classification (n=1) | Determine effects of treatment (n=1) | Identify disease risk / progression (n=6) | Multiple aims (n=37) | Predict disease onset (n=4) | Overall (n=83) |
|---|---|---|---|---|---|---|---|---|---|---|
| **Is paraclinical data being collected?** | | | | | | | | | | |
| No | 3 (75.0%) | 1 (25.0%) | 0 (0%) | 2 (66.7%) | 0 (0%) | 0 (0%) | 2 (33.3%) | 12 (32.4%) | 0 (0%) | 21 (25.3%) |
| Yes | 0 (0%) | 1 (25.0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 2 (33.3%) | 9 (24.3%) | 2 (50.0%) | 15 (18.1%) |
| Missing | 1 (25.0%) | 2 (50.0%) | 1 (100%) | 1 (33.3%) | 1 (100%) | 1 (100%) | 2 (33.3%) | 16 (43.2%) | 2 (50.0%) | 47 (56.6%) |

## 32. What disease specific data being collected? (summary)



## 33. Is data on environmental exposure and life-style being collected?

| | Overall (n=83) |
|---|---|
| **Is data on environmental exposure and life-style being collected?** | |
| No | 46 (55.4%) |
| Yes | 19 (22.9%) |
| Missing | 18 (21.7%) |



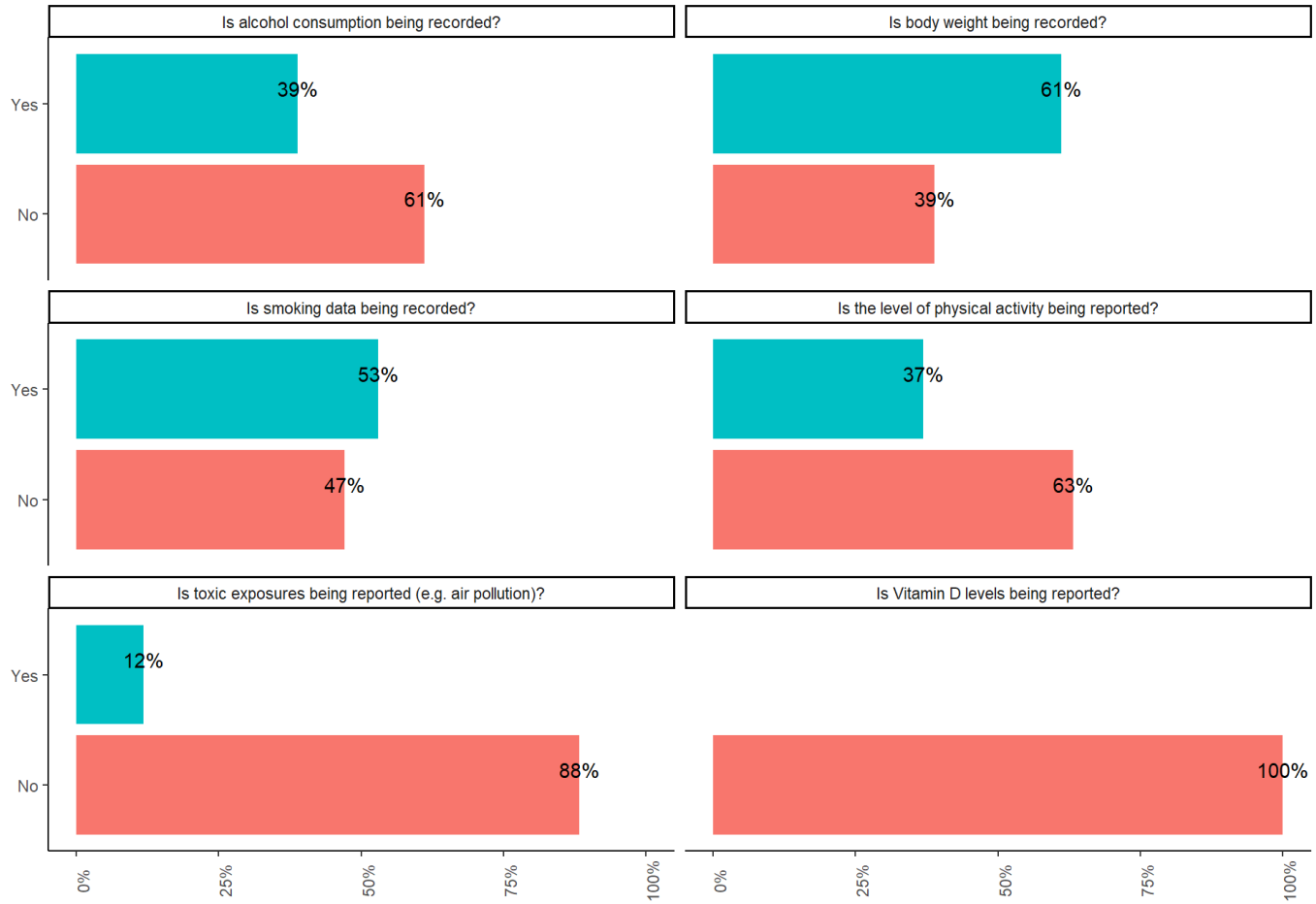
Answers divided by type of study:

| | Biobank (n=4) | Cohort study (n=20) | Consortium study (n=15) | Health administrative data (n=4) | Other (please specify below) (n=12) | Registry (n=9) | Overall (n=83) |
|---|---|---|---|---|---|---|---|
| **Is data on environmental exposure and life-style being collected?** | | | | | | | |
| No | 2 (50.0%) | 10 (50.0%) | 10 (66.7%) | 3 (75.0%) | 11 (91.7%) | 7 (77.8%) | 46 (55.4%) |
| Yes | 1 (25.0%) | 10 (50.0%) | 5 (33.3%) | 1 (25.0%) | 0 (0%) | 2 (22.2%) | 19 (22.9%) |
| Missing | 1 (25.0%) | 0 (0%) | 0 (0%) | 0 (0%) | 1 (8.3%) | 0 (0%) | 18 (21.7%) |

Answers divided by purpose of study:

| | Administrative health care data (n=4) | Determine potential treatment targets, biomarkers (n=4) | Don't know (n=1) | Other (please specify below) (n=3) | Patient/cohort classification (n=1) | Determine effects of treatment (n=1) | Identify disease risk / progression (n=6) | Multiple aims (n=37) | Predict disease onset (n=4) | Overall (n=83) |
|---|---|---|---|---|---|---|---|---|---|---|
| **Is data on environmental exposure and life-style being collected?** | | | | | | | | | | |
| No | 3 (75.0%) | 3 (75.0%) | 1 (100%) | 0 (0%) | 1 (100%) | 1 (100%) | 4 (66.7%) | 28 (75.7%) | 1 (25.0%) | 46 (55.4%) |
| Yes | 1 (25.0%) | 1 (25.0%) | 0 (0%) | 3 (100%) | 0 (0%) | 0 (0%) | 2 (33.3%) | 9 (24.3%) | 3 (75.0%) | 19 (22.9%) |
| Missing | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 18 (21.7%) |

## 34. What environmental exposure is collected?



**35. What environmental exposure is collected (based on several different questions, should probably be % rather than n)?**

# Acknowledgements

The current report corresponds to Deliverable 1.1 “EU-wide mapping report with focus on international databases collections and registries” of Work Package 1 “Data sources and standards for predictions in personalized medicine”.

EU-STANDS4PM is funded by the European Union Horizon2020 framework programme of the European Commission Directorate-General for Research and Innovation under Grant Agreement # 825843.

**CONTACT**

Marc Kirschner (Coordination)
Forschungszentrum Jülich GmbH
Project Management Jülich (PtJ)
52425 Jülich, Germany
Phone: +49 2461 61-6863
E-mail: m.kirschner@fz-juelich.de
www.eu-stands4pm.eu

Sylvia Krobitsch
Forschungszentrum Jülich GmbH
Project Management Jülich (PtJ)
52425 Jülich, Germany
Phone: + 49 30 20199-3403
E-mail: s.krobitsch@fz-juelich.de