

The importance of health data quality for trustworthy AI development and safe AI use

Dipak Kalra
Jens Declerck, Sonia Priou, Sofia Palmieri

The threats to health system sustainability and resilience



Economic context

- ✓ Legacy of crises: public finance deficits
- ✓ Continued increases in public sector health spending anticipated
- ✓ Concerns about how this will be paid for (sustainability of public finances)



Population health

- ✓ Ageing and rising levels of chronic disease and comorbidity
- ✓ Public health challenges
- ✓ Inequalities



Health systems

- ✓ Challenge of responding to changing population needs and demands
- ✓ **Marked variation in clinical practices and outcomes**
- ✓ **Need for structural reforms – digital transformation**

The urgency to deliver health systems value

Increasing value-for-money of health services must be even more strongly emphasised.

Achieving bold efficiency gains by cutting ineffective and wasteful spending, while also reaping the benefits of technology and the digital transformation of health systems, including Artificial Intelligence (AI), is imperative.

OECD, January 2024

Poor Usability of Electronic Health Records Can Lead to Drug Errors, Jeopardizing Pediatric Patients

Challenges can stem from product design, clinician use, and customization

ISSUE BRIEF | April 24, 2019 | Read time: 10 min

Projects: [Health Information Technology](#)

State finds hundreds of medication errors linked to healthcare technology

The majority of errors were attributed to the human-computer interface, workflow and communication, and clinical content, Pennsylvania Patient Safety Authority says.

By [Bill Siwicki](#) | April 10, 2017 | 03:36 PM



Poor Data Quality, Weak Algorithms Lead to Patient Matching Issues

Patient matching issues are exacerbated by poor data quality, insufficient algorithms, and a lack of technology.

Covid: Man offered vaccine after error lists him as 6.2cm tall

🕒 18 February 2021

Leeds Hospital's 'Own Data' Stopped Surgery

The NHS chief who halted children's heart surgery at Leeds General Hospital says the hospital's faulty data was to blame.

🕒 Tuesday 9 April 2013 10:33, UK



Healthcare Weekly Staff

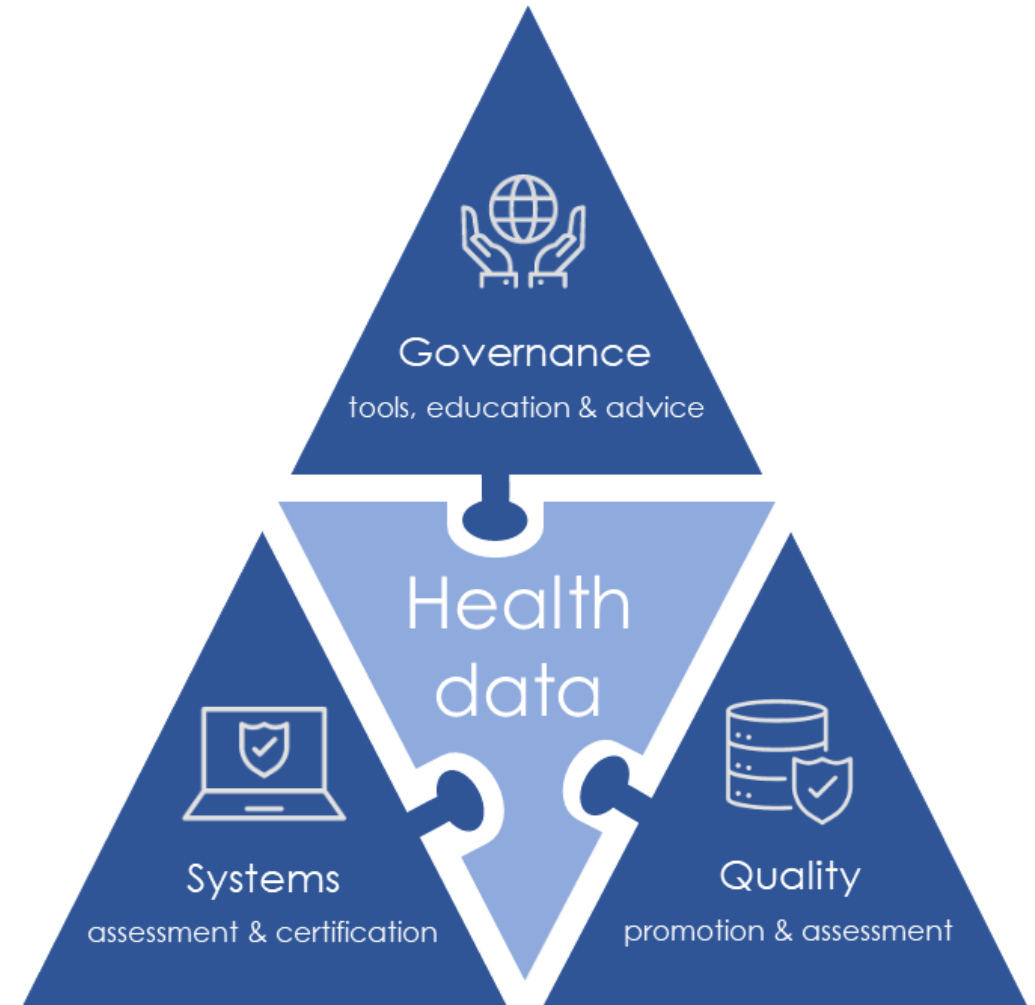
📅 July 16, 2021 👁 341 Views

Healthcare Data Quality Issues Plague Several Hospitals – Are They Preventable?

What is data quality?

- “Fitness for use”
- Multi-dimensional concept
- Multi-stakeholder perspective

*Patients and clinicians
Healthcare managers
Public health agencies
Pharma and industry
Regulators and HTA agencies
Healthcare funders
...*



i~HD holistic view

Data quality dimensions

Completeness

Data values are present

Consistency

Data satisfy constraints

Uniqueness

Patient records are not duplicated

Stability

Data are comparable among sources and over time

Contextualisation

Data are annotated with acquisition context

Representativeness

Data are representative of population

Correctness

Values are true and unbiased

Trustworthiness

Data can be trusted based on owner's reputation

Timeliness

Data is promptly processed and up-to-date

Completeness

- Data items that are known to be collected/accessible, check whether individual data values are complete across patient visit records
- Frequency of missing data per variable:
 - Occurrences of absence (blanks)
 - Occurrences of nonsense (data entered in an incorrect format)
- Completeness score per variable

Completeness
Data values are
present

Consistency

- Consistency by type
 - Examine whether all data values are in the right format, as defined in the data dictionary
- Consistency by range
 - Examine whether numerical values fall within pre-specified ranges and whether categorical/character variables have values that comply with predefined response options as described in the data dictionary
- Consistency by multivariate rule
 - Examine for violations to data quality interdependency relationships that have been defined between different variables

Consistency
Data satisfy
constraints

Correctness

- Assess correctness of a subset of data variables by combining information across variables (multivariate correctness) or over time (longitudinal correctness).
- For example (based on diabetes data set)
BMI based on height and weight

Correctness

Values are true and unbiased

Uniqueness

- Records representing a single patient are not duplicated
 - Number of completely duplicated data rows will be identified
 - Patient records/datasets will be checked to look for identical visit identifiers even though values or one or more data items might have different values
 - Patient records/datasets will be checked for those that had identical data while the visit identifier differed.

Uniqueness

Patient records are not duplicated

Example DQ rule for height

Variable Name	Height	
Definition	Indicate the height of the patient	
Supporting definition	NA	
Inclusion criteria	All patients	
Timing	At index event for CVD	
Data source	Unstructured or semi-structured data extracted from free text through NLP or if available, from structured databases	
Type	Numerical	
Response options	cm	
Data quality rules	Completeness	Data items that are known to be collected/accessible, check whether individual data values are complete across patient visit records.
	Consistency by type	Check whether the format complies with the one specified in the data dictionary.
	Consistency by range	135 – 230 cm
	Correctness	Height vs. Weight → BMI Sensible range: 25 – 45 kg/m ²

Example completeness rules

Completeness

- For all variables, except for:
 - Use of insulin pump; Can only be complete if 'types of diabetes treatment' equals insulin.
 - Glycaemic control – mean glucose; Can only be complete if sensor-based continuous glucose monitoring equals yes.
 - Glycaemic control – SD of mean glucose; Can only be complete if sensor-based continuous glucose monitoring equals yes.
 - Glycaemic control – TIR; Can only be complete if sensor-based continuous glucose monitoring equals yes.
 - Glycaemic control – TIH; Can only be complete if sensor-based continuous glucose monitoring equals yes.



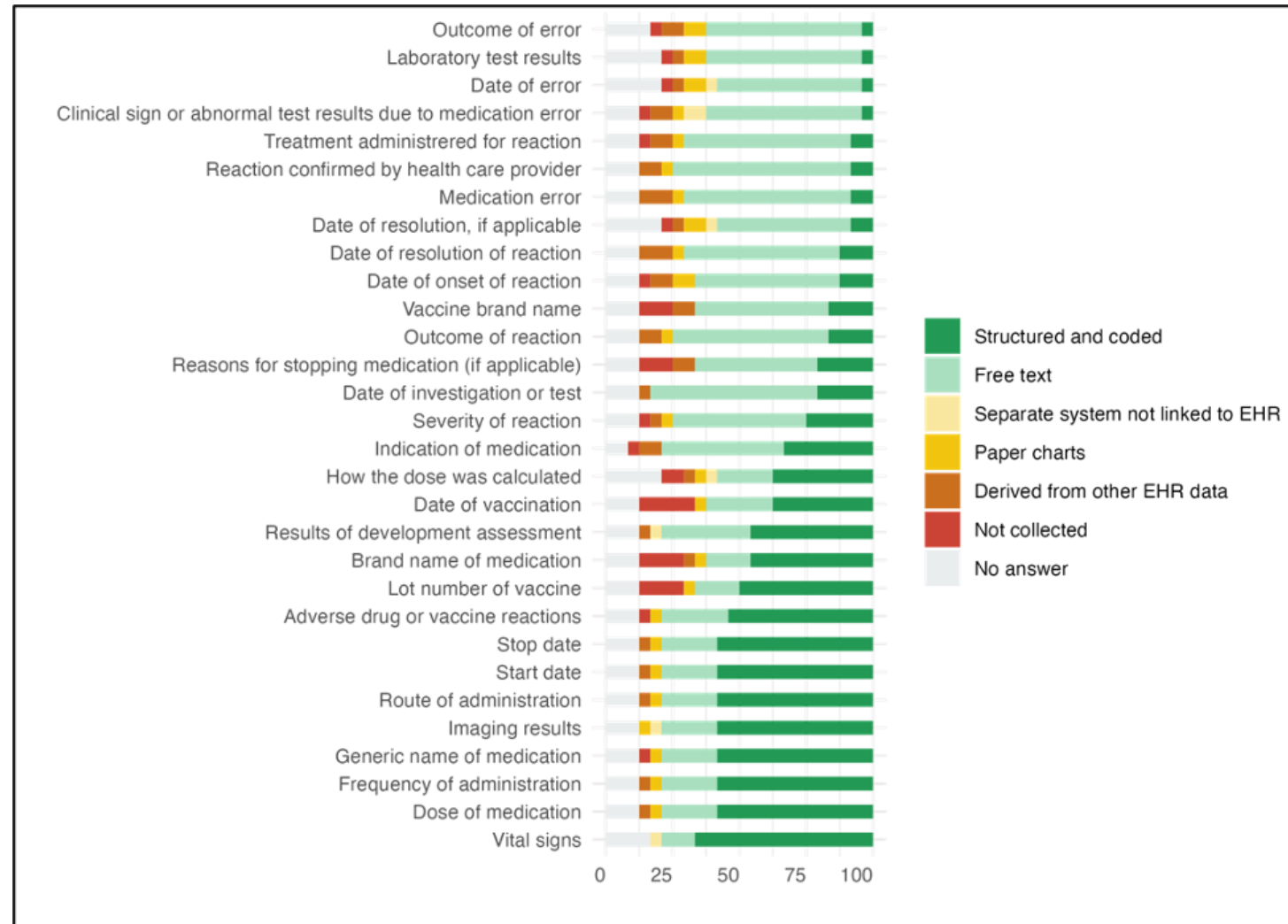
Example consistency rules

Consistency (by range)

- By range: The following ranges have been pre-defined for numerical variables, and their consistency (by range) should be checked against them:
 - HbA1c; 4,0 – 6,0 rel. %
 - Mean of glucose; 40 to 500 mg/dL
 - SD of mean glucose; 0 – 350 mg/dL
 - TIR; 0 – 100%
 - TIH; 0 – 100%
 - Total cholesterol; < 200 mg/dL
 - LDL cholesterol
 - Target value at low risk < 116 mg/dL
 - Target value at moderately increased risk < 100 mg/dL
 - Target value at high risk < 70 mg/dL
 - Target value at very high risk < 55 mg/dL
 - HDL cholesterol
 - Male > 55 mg/dL
 - Female > 65 mg/dL
 - Triglycerides; < 150 mg/dL



The availability of structured information on medication safety and vaccines, in children

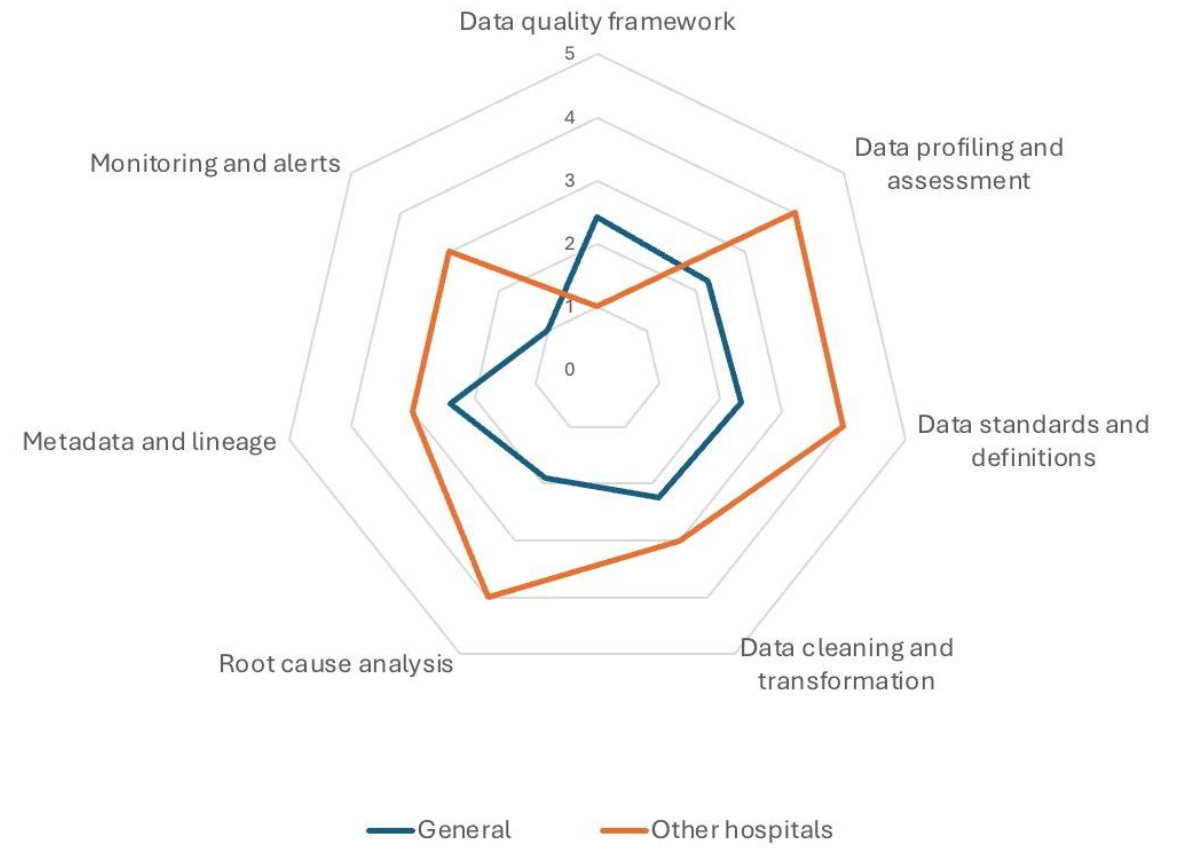
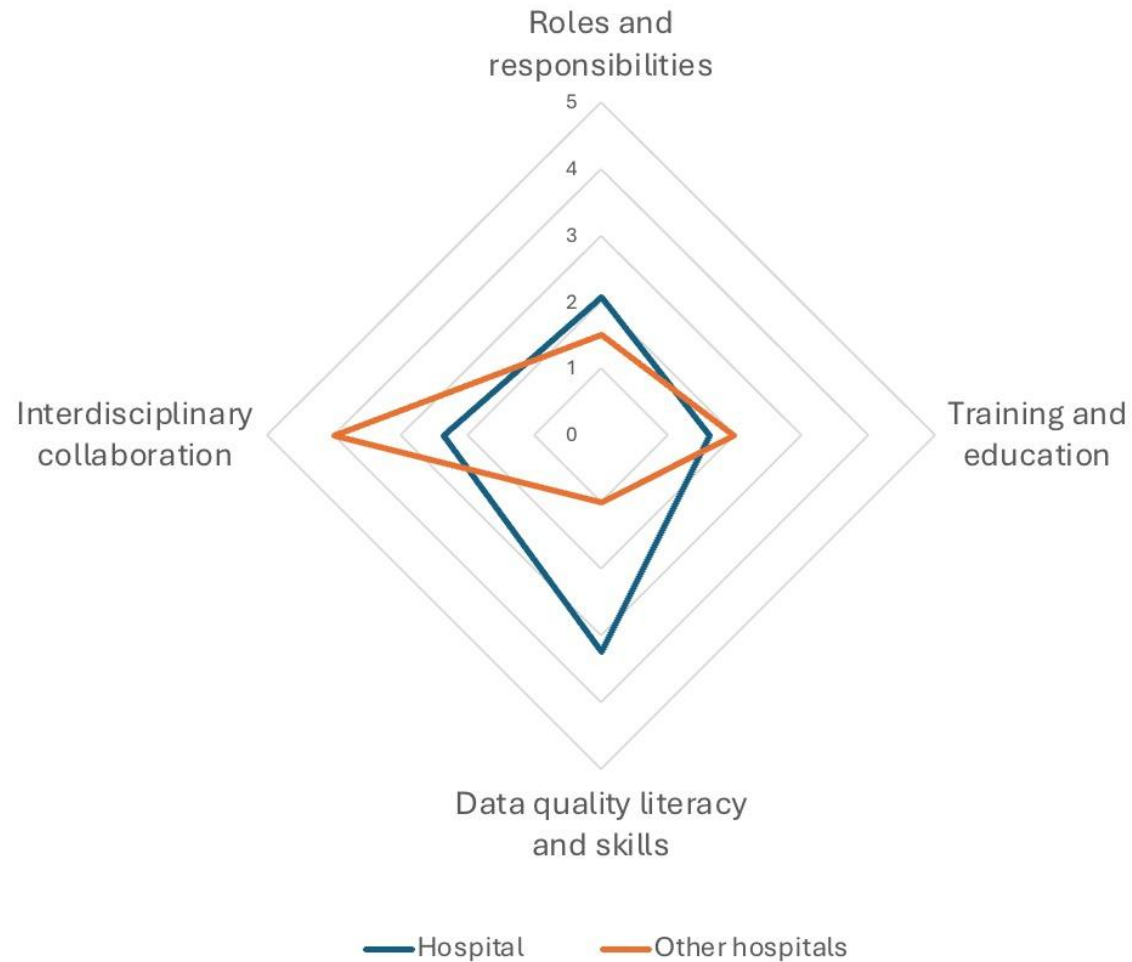


Drug and Vaccine Safety data in the EHR

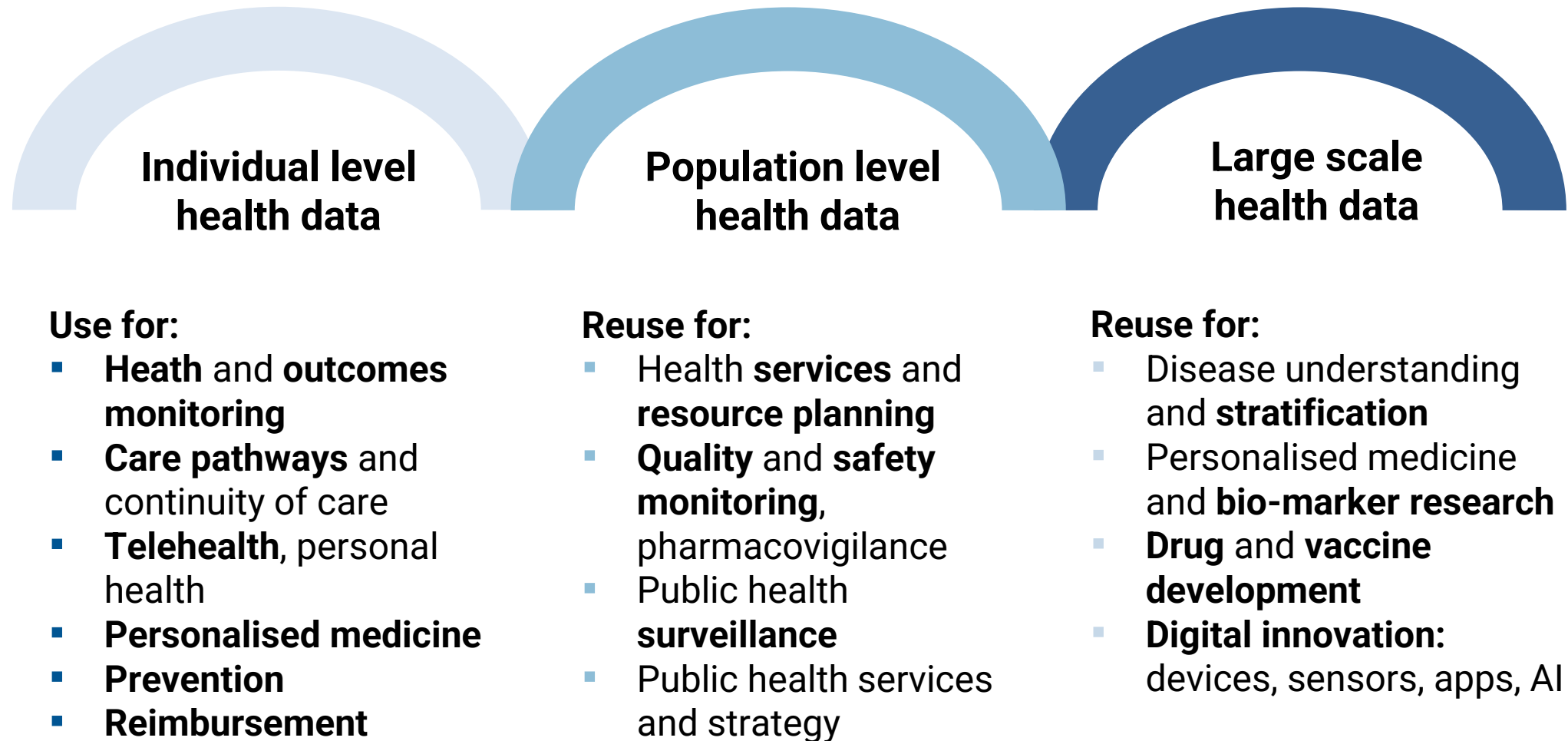
Each colour represents the percentage of sites collecting the data item in a particular format
(n = 24 hospitals across Europe, collected during 2024)

<https://doi.org/10.2196/72573>

Providing feedback to hospitals about their data quality maturity

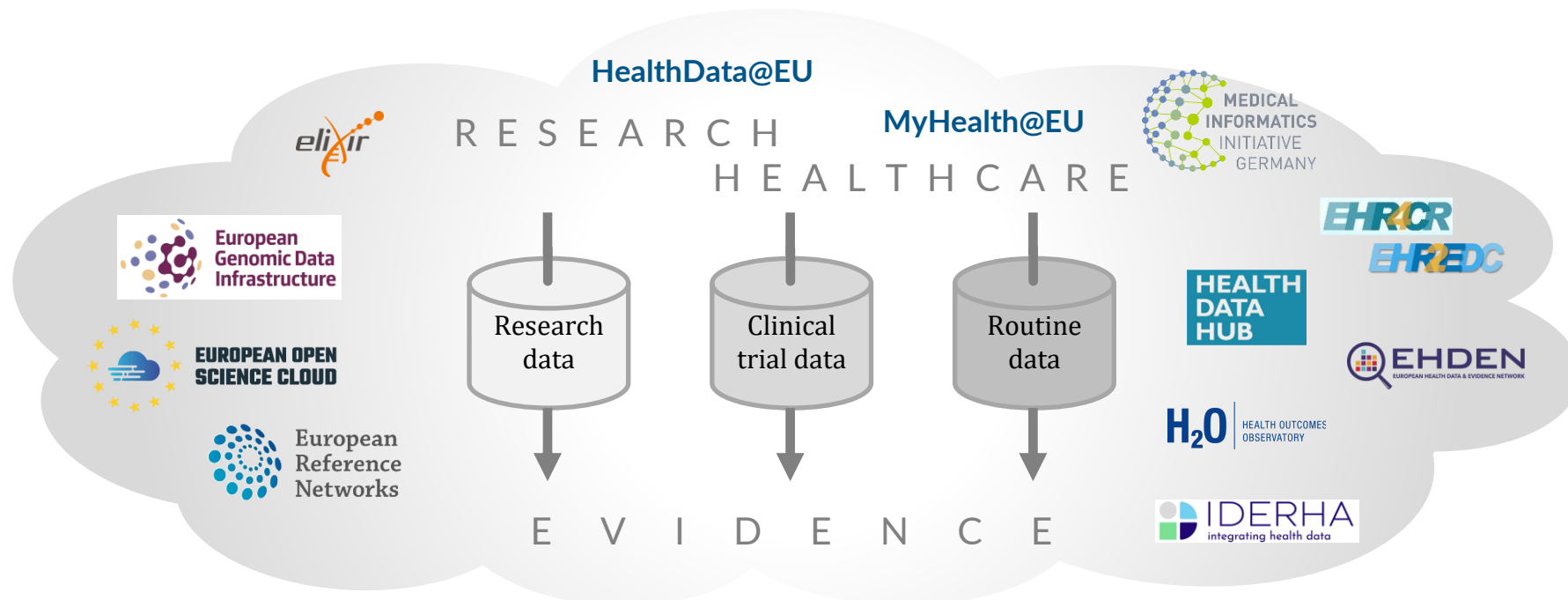


Massive opportunities to learn from health data, at all scales

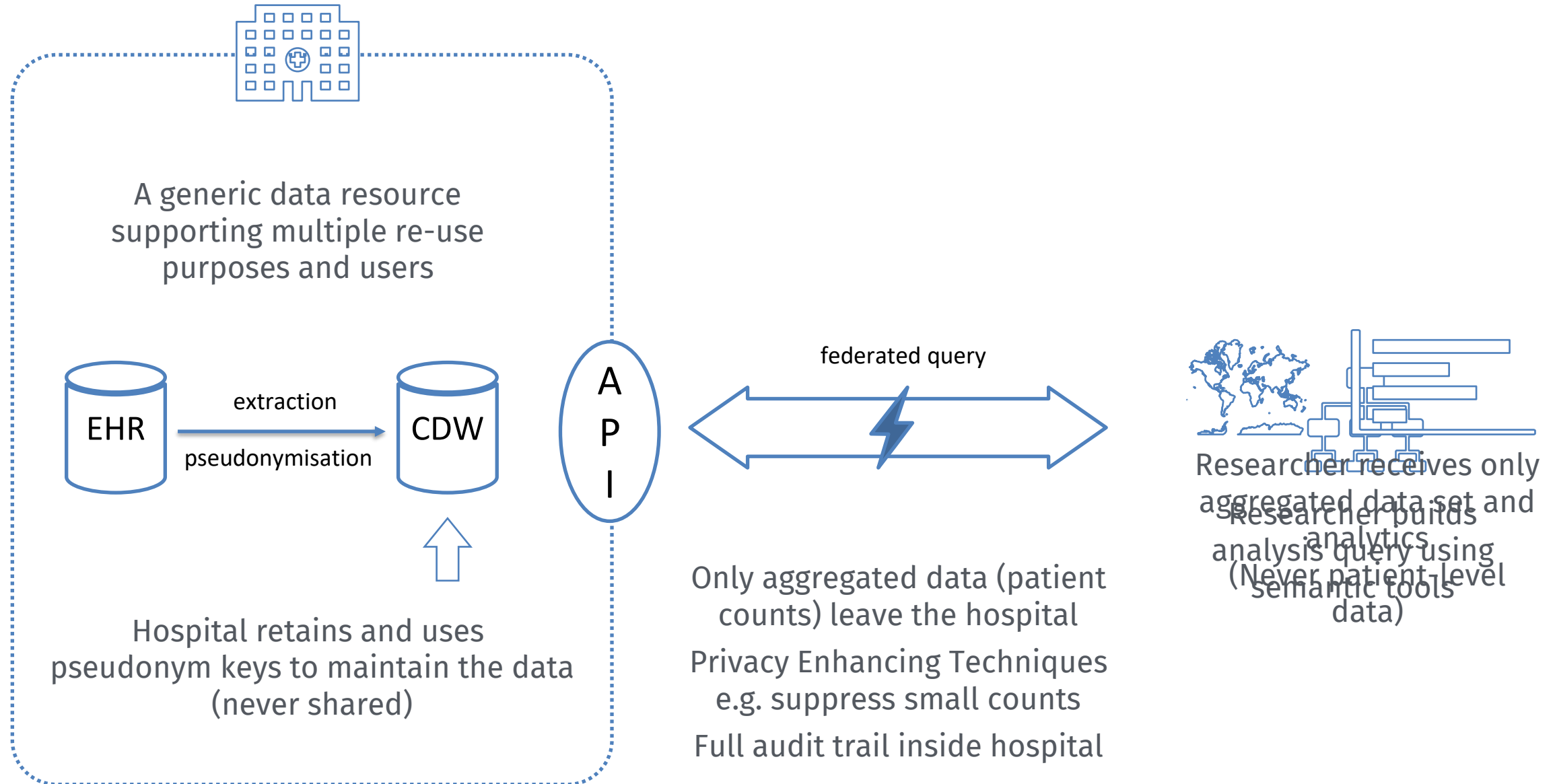


Big health data sharing initiatives

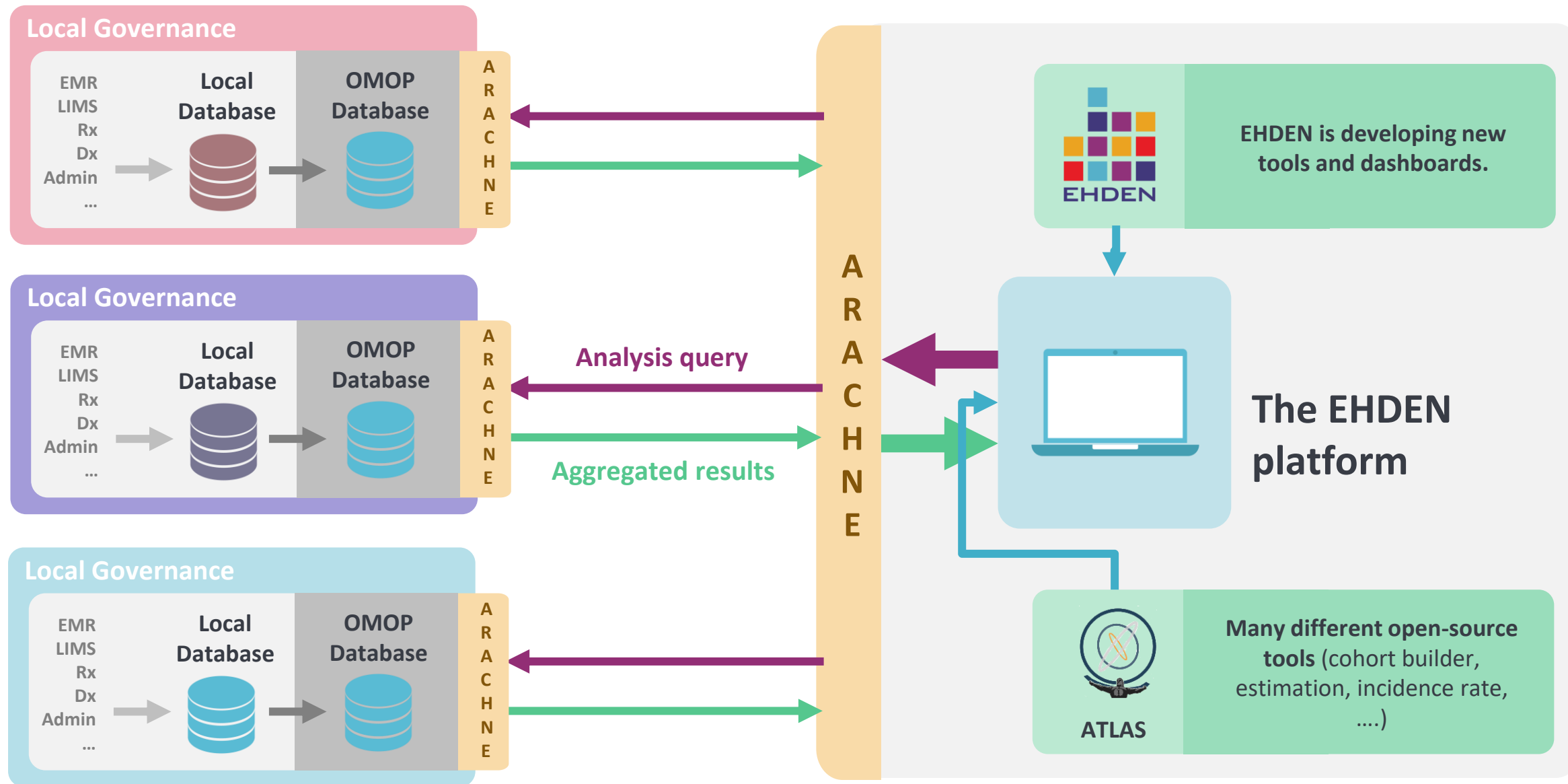
- Multiple initiatives are scaling up health data access
 - across jurisdictional, institutional and domain borders, for care or for research
- Emerging paradigm for analysing personally-identifiable health data:
 - federated infrastructure model: network of repositories with an overarching governance and interoperability layer



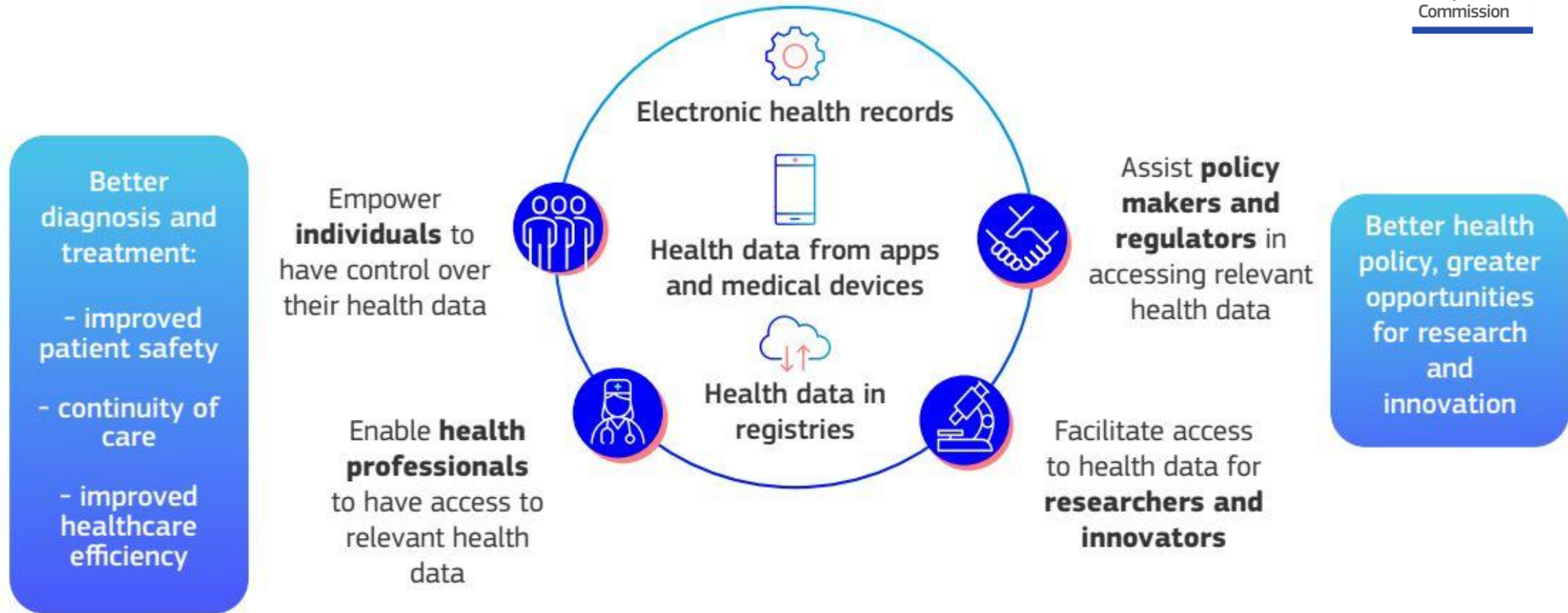
The federated query data flow



THE EHDEN FEDERATED DATA NETWORK



The European Health Data Space



Types of data

- Health, healthcare, broad determinants of health
- Socio-economic, environmental, education, occupation
- Behavioural health
- Pathogens
- Genetic, genomic, all 'omic and molecular data
- Automatically-generated personal health data
- Health insurance, claims, reimbursements

Sources of data

- EHR
- Medical/in vitro devices
- Wellness apps
- Registry data (many kinds)
- Clinical trials, studies, investigations that have ended under the Clinical Trial Regulation
- Research cohorts, surveys (after first publication)
- Biobanks

Ancillary data

- Treating health professional details
- Aggregated health needs, access to services
- Healthcare financing, resource allocation

EHDS permitted and prohibited purposes for secondary health data use

Public interests for public and occupational health

- cross-border threats to health
- public health surveillance
- healthcare quality and patient safety
- safety of medicines and devices

Policy making and regulatory activities

Statistics related to health and care

Higher education and teaching in health and care

Scientific research contributing to health, HTA or care

- product and service development and innovation (e.g. medicines)
- training, testing and evaluating of algorithms, digital health tools

Improving and optimising delivery of care

Developing products or services that may legally, socially or economically harm individuals or groups

- illicit drugs, alcoholic beverages, tobacco products
- products or services that cause addiction or contravene public order

Decisions with effects detrimental to a person based on their electronic health data

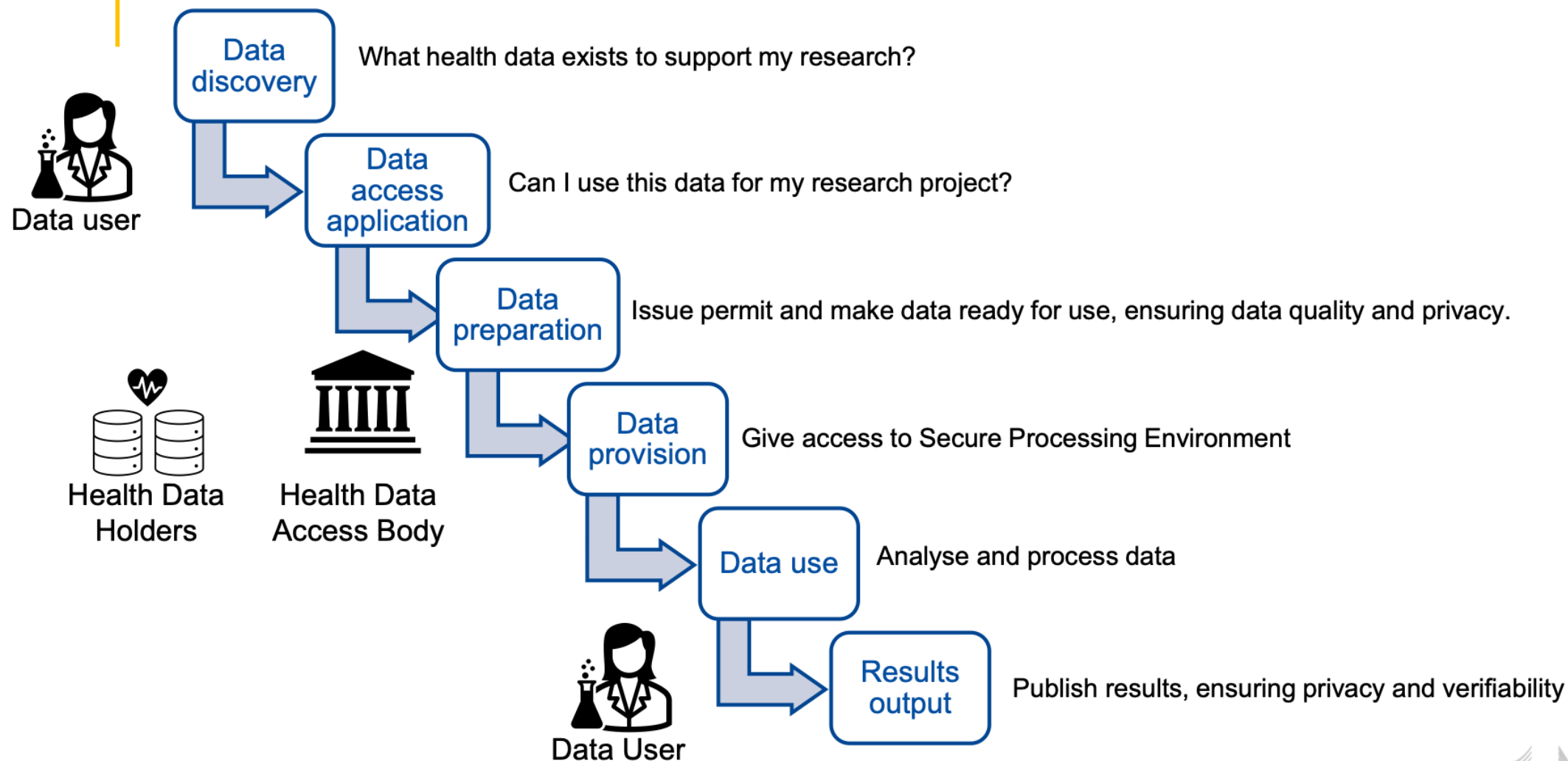
- e.g. legal, employment, insurance, pension, banking, mortgaging of properties

Decisions that exclude persons or groups, or provide less favourable terms, for products and services

Advertising or marketing activities

Activities in conflict with ethical provisions pursuant to national law

User journey



Data catalogue information properties

Standard properties: Theme = HEALTH, Type = PERSONAL_DATA, Access rights = NON_PUBLIC

Always provided

- Unique identifier
- Category e.g. EHR, biobank
- Title and description
- Descriptive keywords
- Health themes (coded)
- Applicable legislation
- Landing web page
- Publisher and contact
- Provenance (origin, creation)
- Purpose of creation
- Geographic coverage
- Sample data extract or mockup

Recommended

- Schema conformity
- Coding systems e.g. SNOMED and actual codes used
- Legal basis for the data set
- Data privacy ontology terms
- Data Quality and Utility label
- Accrual periodicity
- Publications used or referenced
- Source, if a derived data set
- Related data sets

Recommended

- Population demographic coverage
- Time period covered by the data
- Number of records
- Number of individuals
- Age ranges
- Links to any data distribution analytics

Plus other optional properties

Application procedures

Data Access Application

For processing personal electronic health data.

Includes **detailed application requirements** such as applicant details, data description, intended use, ethical assessments (where required by MS law), and security measures.

Data can be accessed in **pseudonymised format** unless **anonymised data** suffice for the purpose.

If accepted, the applicant receives a **data permit**.

= **direct access to data under stricter conditions.**

Data permits are generally granted for up to 10 years with possible extensions.

A streamlined procedure across EU

Single application form

Single permit template

Fees based on the complexity and duration of data access.

Data Request

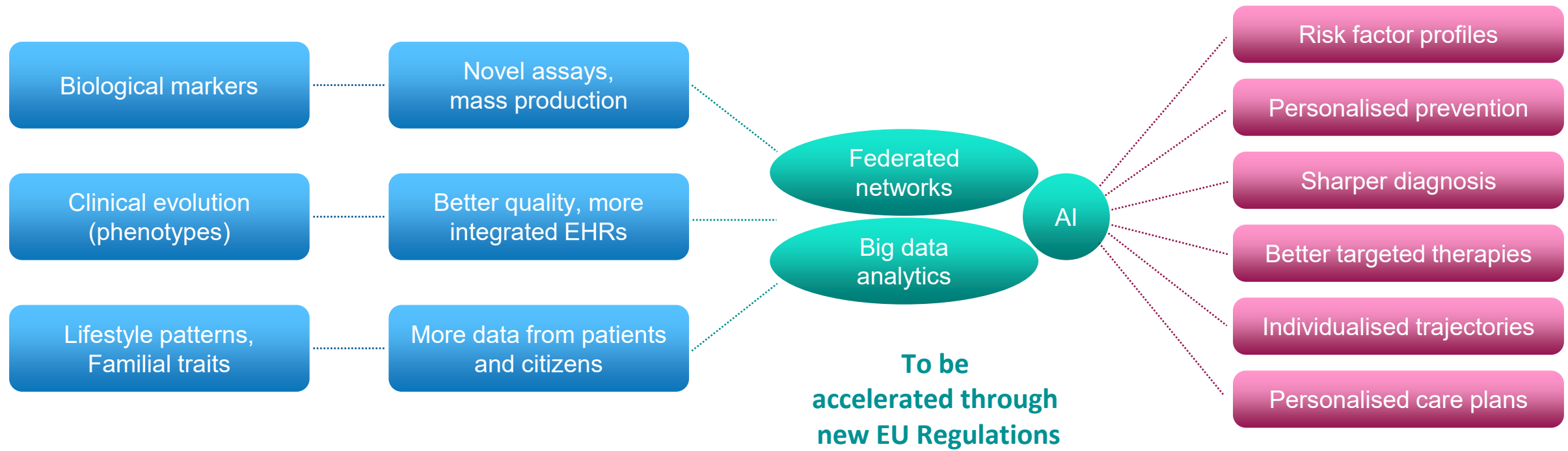
For obtaining answers in **anonymised statistical format** only.

Less stringent application focused on identity, intended use, and safeguards without direct access to personal data.

= **only statistical outputs from anonymised data, suitable for broader or public interest inquiries without personal data access.**

Secure Processing Environment (SPE)

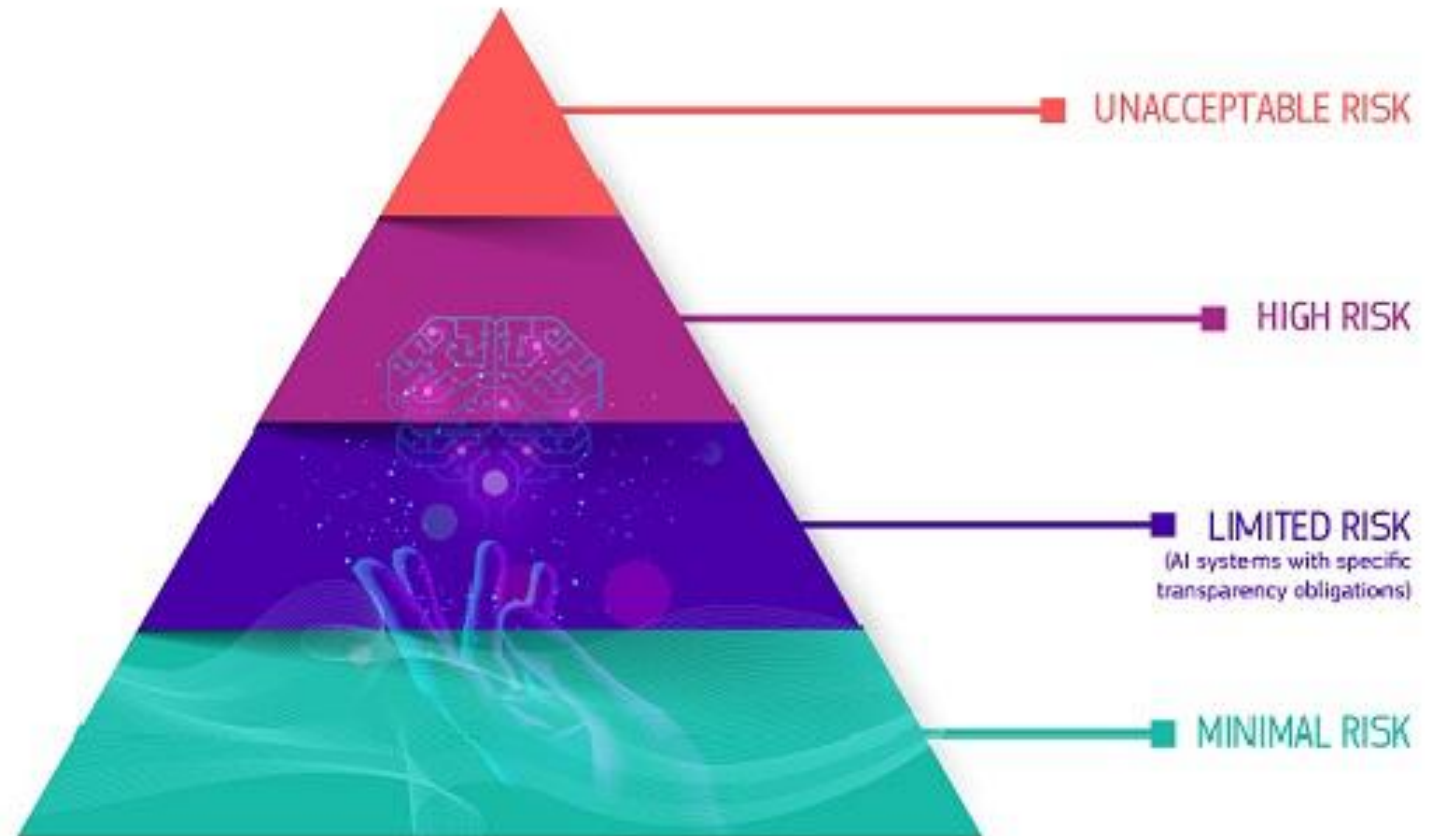
- SPE setup to restrict data access to authorised users
- Access limited to data adequate, relevant, and necessary for specific, approved purposes
- Pseudonymised data provided unless anonymised data suffices, with strict controls on de-identification
- Download of personal data strictly prohibited
- State-of-the-art measures to prevent unauthorised data modification, access, or removal
- Logging and monitoring of activities within the SPE for compliance and audit purposes



- Utilise the history, examination findings, labs, echo and electro cardiography to quantify the probability
 - that the patient has a diagnosis, or is at high risk of developing, HCM (e.g. young athletes)
 - that the patient will develop an arrhythmia in the future
 - of HCM occurring within the next 1-3 years,
 - of any patient-modifiable risk factors e.g. lifestyle
 - if generalist should refer a patient to an HCM specialist
 - of sudden cardiac death caused by an arrhythmia
 - estimate the risk of progressive hypertrophy causing obstructive symptoms and potentially leading to heart failure

“In the health sector where the stakes for life and health are particularly high, increasingly sophisticated diagnostics systems and systems supporting human decisions should be reliable and accurate.

High-risk AI systems shall be designed and developed in such a way to ensure that their operation is sufficiently transparent to enable users to interpret the systems’ output and use it appropriately. An appropriate type and degree of transparency shall be ensured.”






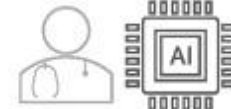
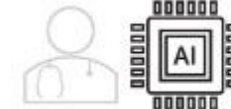
Obligations for high risk AI

- **Adequate risk assessment and mitigation systems**
- **High quality datasets feeding the system to minimise risks and discriminatory outcomes**
- Logging of activity to ensure traceability of results
- Detailed documentation providing all information necessary on the system and its purpose for authorities to assess its compliance
- Clear and adequate information to the user
- Appropriate human oversight measures to minimise risk
- High level of robustness, cybersecurity and accuracy
- Conformity assessment (by notified body) and EU registration

Ensuring good health data for the training of AI algorithms

1. Perform a risk assessment, plan mitigation measures
2. Appoint a data quality manager, set up a data quality process
3. Select appropriate data sources to the intended context of use
4. Select training and validation data covering appropriate patients
5. Assess data quality, representativeness, mitigate bias
6. Document the use of data (transparency, explainability)

Assessing risk in relation to AI autonomy

Assistive AI algorithms			Autonomous AI algorithms		
	Level 1	Level 2	Level 3	Level 4	Level 5
	 Data presentation	 Clinical decision-support	 Conditional automation	 High automation	 Full automation
Event monitoring	AI	AI	AI	AI	AI
Response execution	Clinician	Clinician and AI	AI	AI	AI
Fallback	Not applicable	Clinician	AI, with a backup clinician available at AI request	AI	AI
Domain, system, and population specificity	Low	Low	Low	Low	High
Liability	Clinician	Clinician	Case dependent	AI developer	AI developer
Example	AI analyses mammogram and highlights high-risk regions	AI analyses mammogram and provides risk score that is interpreted by clinician	AI analyses mammogram and makes recommendation for biopsy, with a clinician always available as backup	AI analyses mammogram and makes biopsy recommendation, without a clinician available as backup	Same as level 4, but intended for use in all populations and systems

Bitterman, Danielle S et al. Approaching autonomy in medical artificial intelligence. The Lancet Digital Health, Volume 2, Issue 9, e447 - e449

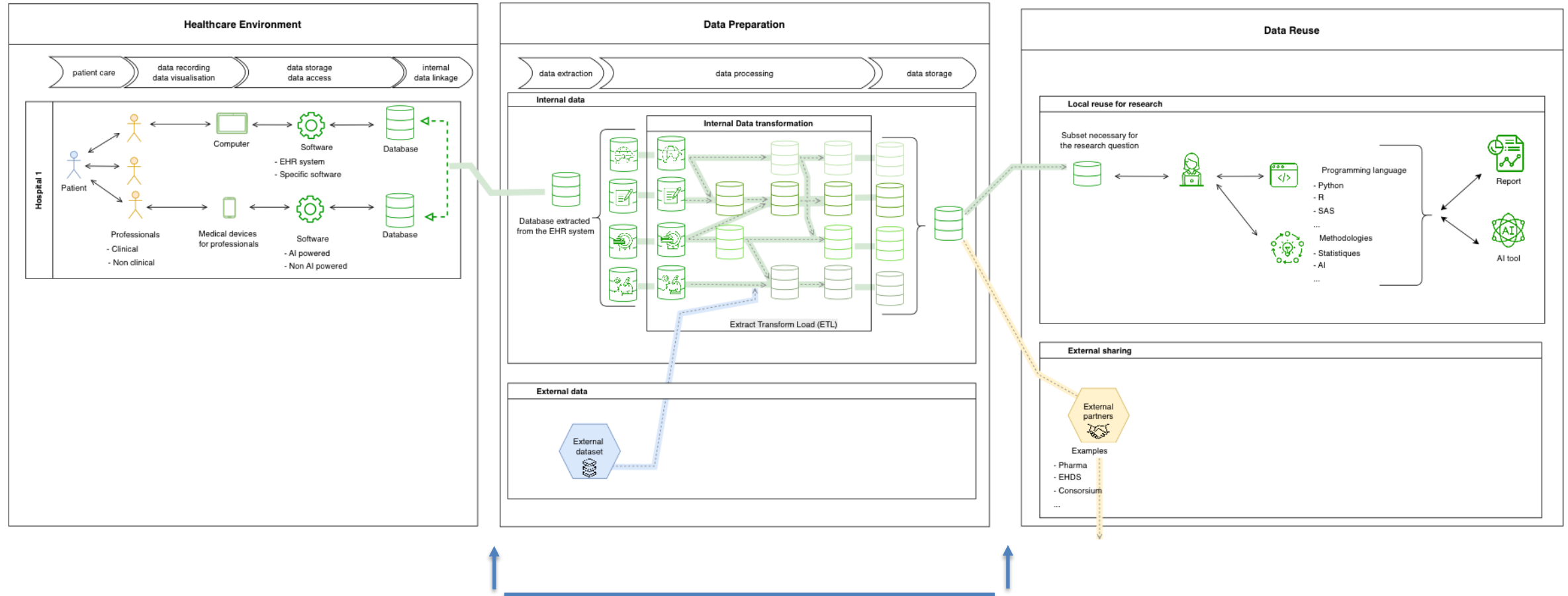
Data must be relevant, high-quality, and representative of the patient populations to be served

- Data used to train and validate healthcare AI systems must be **examined for potential biases**, such as overrepresentation of certain ethnicities, genders, or age groups, and underrepresentation of others
- Developers are expected to **document their data sources, assess their suitability** for the intended purpose, and describe the steps taken to identify and mitigate biases
- There is also an expectation of **ongoing monitoring**, particularly for systems that learn and adapt over time

Appoint a Data Quality Manager to:

- Develop a Standard Operating Procedure, proportionate to the risk assessment
- Define and oversee the data quality tools and processes to be used, ensure users are trained and tools are kept up to date
- Establish and maintain a library of dimensional data quality rules
- Conduct continuous and discrete data quality assessments for both training and validation data
- Identify and investigate data quality errors when they are detected.
- Collaborate with all stakeholders within the organisation and with data sources to implement corrective actions when data quality errors are found
 - when to apply statistical corrections
 - when to use synthetic data to compensate for bias
 - when to reject data as being unsuitable
- Formalise how data quality and bias assessments and mitigations are reported in system documentation, transparency notices, regulatory submissions
- Ensure that all members of the organisation are appropriately trained

Tracking complex data mapping pipelines = quality risks or opportunities



For example, compare the quality profile at these two points:
has it degraded or been enhanced?

About the AI system’s intended deployment and use

Countries	
Healthcare settings	
Patient populations	
Disease areas	
Care pathway scenarios	
Care decisions to be supported	

Data set name	
Data set size (number of patients included)	
For retrospective (real world) data source:	
Country	
Data provider organisation type: e.g. hospital, GP, registry, claims, other	
Data currency data range	
Data ingestion process: e.g. manual transcription, direct import, structural and terminology mappings, NLP, data cleaning, other	
For synthetic data source:	
Profile of the real-world data used for SD generation	

For prospectively collected data and/or a validation study:	
Country	
Healthcare organisation type: e.g. hospital, GP, other	
Recruitment methodology	
Intervention using the AI system	
Devices used	
Sample size	

Characteristic	Value distribution in the data set ¹	Is this an inclusion or exclusion criterion?	Target class balance ratio ²
Age distribution			
Gender distribution			
Race, ethnicity and cultural aspects			
Lifestyle factors and socio-economic status			
Main condition(s)			
Condition name or disease area			
Longevity of the condition			
Severity of the condition			
Disease trajectory lifecycle points			
Comorbidity patterns			
Standards of care being used e.g. prevalent clinical guideline(s)			
Medication usage pattern			
Other patient characteristics			

1. This should be provided using statistical quantitative metrics, e.g. age: median and interquartile range; condition name as a coded clinical term(s); severity using an ordinal scale or coded clinical term(s). Where there are established metrics for a data element these should be used if they apply.

2. This column should be used to indicate the alignment of the data set value distributions with published population distributions regarding the patients, conditions and treatments in the data. The source of the published reference data should be provided, and how the comparison has been made.

Evidencing that high quality data has been used

Data element	Dimensions * and rules	Data lifecycle point **	Minimum threshold	acceptability

* The data quality dimensions that could be used are:
Completeness, Consistency, Correctness, Timeliness, Stability,
Contextualisation, Representation, Trustworthiness, Uniqueness

** Data lifecycle point (at the source, after ingest from the
source, after transformation to a common data model and
semantics etc.)

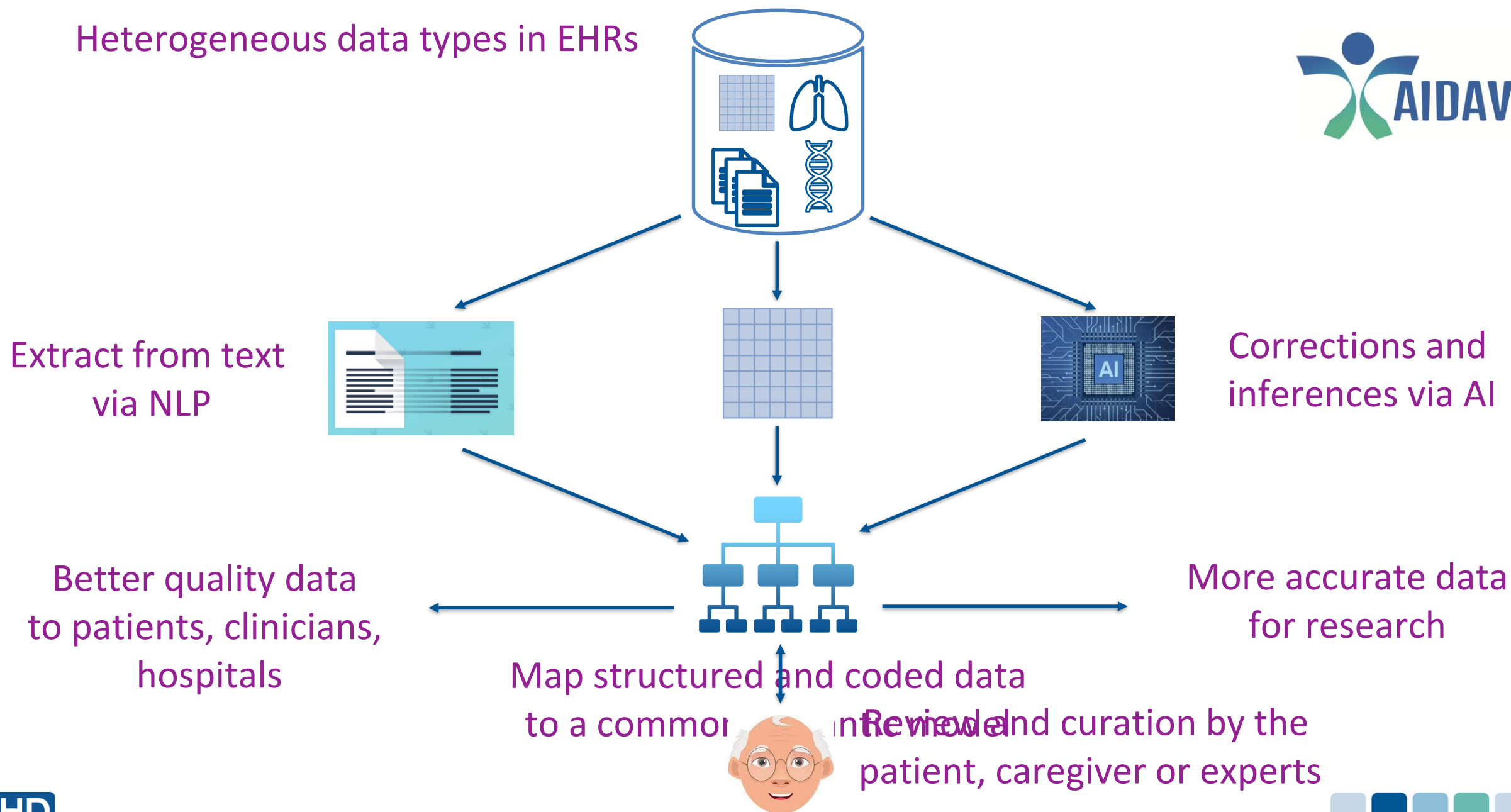
Dimension	Quality assessment	Conclusion about the quality	Decision/action *
Data element name:			
Data element name:			
Data element name:			

* Decision/action: reject the data element, clean, impute missing values...

Enhancing the quality of already-collected data



Heterogeneous data types in EHRs

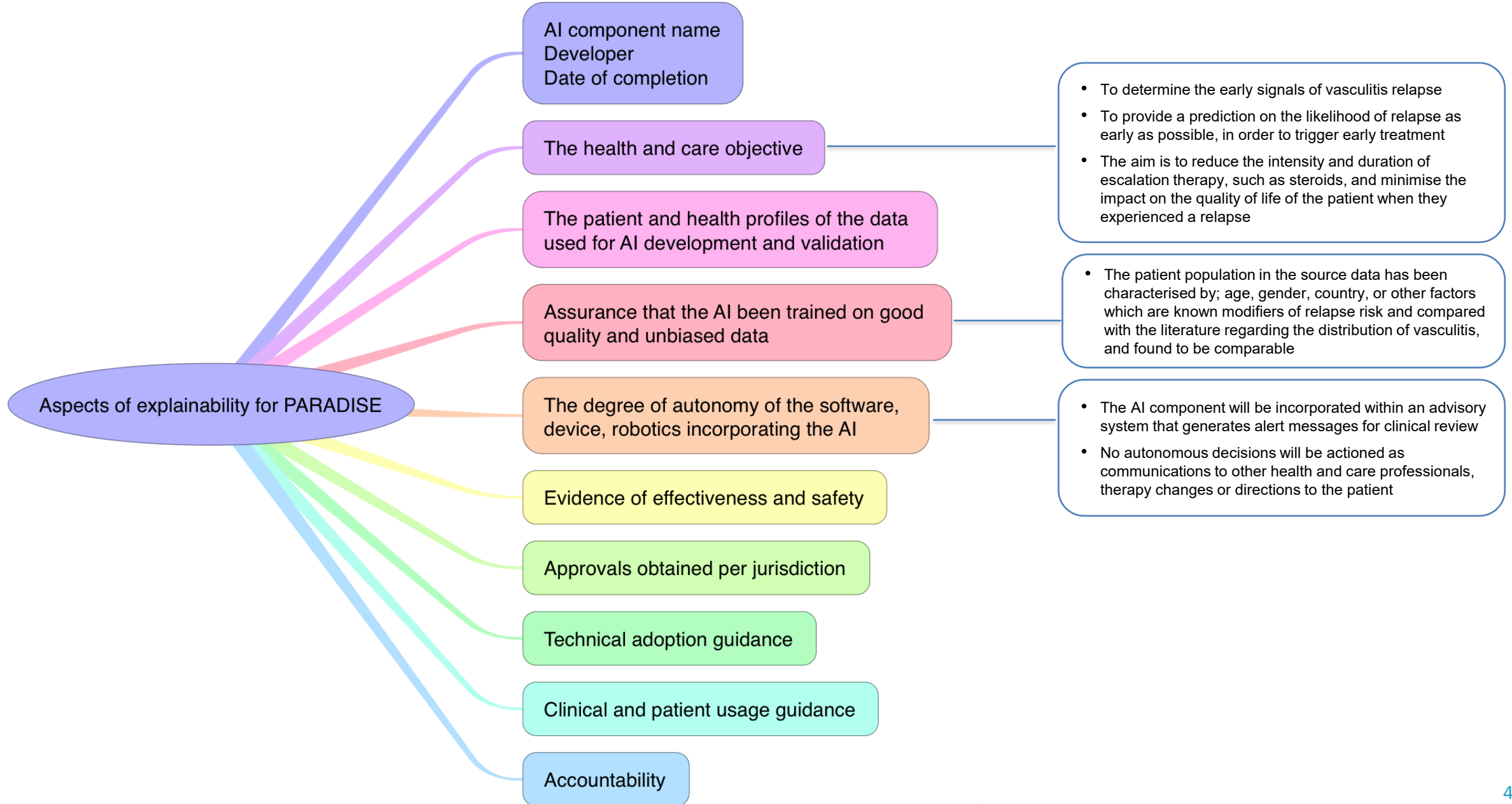


Obligations for high risk AI

- Adequate risk assessment and mitigation systems
- High quality datasets feeding the system to minimise risks and discriminatory outcomes
- Logging of activity to ensure traceability of results
- Detailed documentation providing all information necessary on the system and its purpose for authorities to assess its compliance
- **Clear and adequate information to the user**
- Appropriate human oversight measures to minimise risk
- High level of robustness, cybersecurity and accuracy
- Conformity assessment (by notified body) and EU registration

What information will give patients confidence to trust AI?

- When their clinician is using AI to
 - confirm their diagnosis
 - predict their risks of deterioration or complication
 - determine the most suitable treatment
- When they are invited to use an AI device or app to
 - monitor their condition at home and track when there is a concern
 - advise on drug dosing or a lifestyle risk factor
 - escalate care when it is appropriate



AI component name

Developer

Date of completion

The health and care objective

The health condition being addressed, the challenging clinical or patient decision the AI helps with, including if its intended role is diagnostic, risk or care pathway stratification, personalisation of treatment, early detection of the need for care escalation etc.

Description of the patient and health profiles of the data used for AI development and validation

This should largely dictate the scope of patient populations on whom there is likely to be reliable evidence of its safety and effectiveness, such as the age range, ethnicity, geography, health condition(s), severity, kinds of treatment included etc.

Quality and bias assessments performed on the AI training data, and any corrections applied

How quality, bias and representativeness (equity) have been assessed and what mitigations and corrections have been applied (or recommended limitations of use) to compensate for biases that could not be eliminated.

The degree of autonomy of the software, device, robotics incorporating the AI

If the implemented component is providing advice to the clinician or patient, issuing an alert or warning, taking an action or controlling an instrument such as a medication delivery closed loop system, and if its advice is normally going to be co-interpreted with other decision influencing information that a clinician will utilise in order to arrive at a final decision.

Approvals obtained per jurisdiction

This may include European level such as EU Medical Device Regulation certification and AI Regulation certification, and national level such as HTA approvals.

Evidence of effectiveness and safety

What evidence has so far been accumulated about patient safety, clinical effectiveness, impact on patient outcomes and health economic value.

Technical adoption guidance

Clear guidance to healthcare organisations about how to install and connect the AI containing solution including what input data flows (e.g. EHR data) it will require to perform its reasoning, the format of its outputs and how these may be audit logged and persisted by the adopting organisation, and what data flows are needed back to the developer to continue the machine learning cycles.

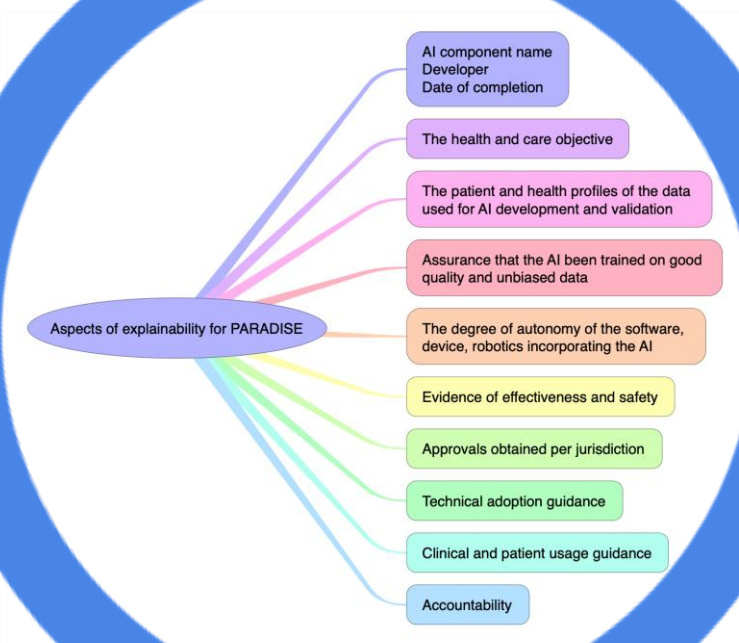
Clinical and patient usage guidance

Clear guidance to healthcare professionals about how to use the AI solution within care pathways, what background competences are needed and what training should be provided, and/or how to train and guide a patient user if applicable, when it is appropriate to use the AI and when not to, scenarios in which system error messages might be generated and what to do, when to over-ride the AI output (if it has some degree of autonomy).

Accountability

Where liability and accountability lie when users follow AI advice or give it serious weight in their decision-making but the advice proves to have been incorrect, or conversely what liability would exist for users choosing not to follow AI advice if it subsequently transpires that it would have been correct.

Mixing the ingredients for the AI clinician factsheet



- Other factsheets: i.e. Coalition for Health AI (CHAI)-Model card
- AI ACT requirements over information
- MDR requirements over information
- HCP friendly



HCP factsheet content

The health and care objective,

• Data used for development and validation,

• Data Quality & Bias Safeguards,

• AI Autonomy & Supervision Safeguards,

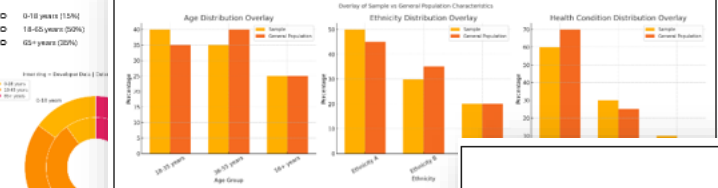
Evidence of effectiveness and safety,

Approval obtained per jurisdiction,

Technical adoption guidance,

Clinical usage guidelines,

• Disclaimers

AI component name Developer Date of completion	
The Health and care objective What is the main goal of this AI in healthcare? Health Condition Addressed: i.e. below Diagnosis (Helps identify diseases or conditions) Risk Stratification (Assesses patient risk levels) Care Pathway Optimization (Guides treatment pathways) Personalization of Treatment (Tailors treatments to individual patients) Early Detection of Deterioration (Identifies need for care escalation) Other:	
The patient and health profile Patient Population Breakdown for AI Validation Age Groups: 0-18 years (15%) 18-65 years (55%) 65+ years (30%) Gender: Male (50%) Female (50%) Ethnicity: White (70%) Black (10%) Asian (10%) Other (10%)	
Overall data representativity 	
Data Quality & Bias Safeguards How was the quality of the data assessed? What findings were made regarding data quality? How was bias assessed in the data? What types of bias were identified? What actions were taken to mitigate biases? What corrections were applied, or what limitations of use were recommended? Overall data quality level (if such thing exist): AI Autonomy & Supervision Safeguards 1. Autonomy level of the AI system 2. Measures to prevent automation bias 3. Alerts or warning systems 4. Practical Interaction, Oversight & Override Mechanisms Evidence of effectiveness and safety Info on the evidence gathered about patient safety, impact on patient outcome and health economic value:	
Approval obtained per jurisdiction MDR certified, HTA approved, AI act certified (link to each certification number) Technical adoption guidance Deployment & Data Requirements 1. Data Input Required 2. Integration models Data Flow & Privacy Management 3. Data Logging & Compliance: 4. Data Privacy & Security Measures: 5. Data Transfer for Improvement: System Maintenance & Updates 6. Software Update Schedule & Process: 7. Predefined System Changes: 8. Expected Lifespan: 9. Contingency Plans for Failures/Downtime: Clinical usage guidelines Intended Users Professional Profile(s): Required Competencies/Training to Use the Tool: How to Use the Tool How to Access/Activate the AI Tool: Steps to Input Data: Expected Output Format:	
i.e. Estimated \$2.3M annual savings at a regional hospital network due to reduced redundant (...) i.e. A 25% drop in 30-day (...) Human Oversight and Decision Triggers When Should the User Engage the AI? ("When to Press the Button"): What to Watch for in the Output (Critical Appraisal Triggers): Situations in which the AI will not run and messages it may provide: Oversight Triggers: When Must a Human Review or Intervene? Discrepancies Between AI and Clinical Practice Guidelines Protocol When AI and CPGs Differ: Escalation Pathway: Overriding AI Recommendations Conditions for Overriding AI Output: How to override: Audit trail and documentation: Disclaimers Various:	

The Digital Patient

Micro-biome

Proteome

The Human Genome

Biobanks

The connected patient

The Quantified Self

Watches

Health kits, Research kits

Trackers

Social media

Simulation

Virtual communities

Mathematical modelling

Living Labs

Real World Data

Virtual Physiological Human

Real Word Evidence

Big Data Spaces

In the last 5 years, more scientific data has been generated than in the entire history of mankind

90% of the data in the world today has been created in the last 2 years

Personal sensor data is expected to grow from 10% of all stored information to ~90% within the next decade

> 1 billion people have access to mobile broadband internet

There are almost as many personal assistant AI bots on the planet as people

Disease registries

Claims databases

Electronic health and care records

Cohort studies and biobanks

Personal health records

Clinical trials data, electronic case report forms

Mobile health apps

Wearable sensors

Geo-sensors

Air quality

Social networks

Lab on a chip

Climate

Food

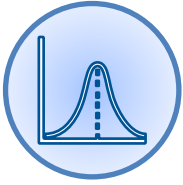
* This is not a comprehensive list

6 grand challenges



Limited uptake of interoperability standards

e.g. reimbursement systems do not favour care collaboration



Prevalence of poor (structured and coded) data quality

e.g. too little data is analysed by healthcare organisations



Data protection concerns often block data sharing and reuse

e.g. insistence that explicit patient consent is required for every data use



Lack of incentives to share data: for care and research

e.g. data is a secret sauce and not a common resource



Digital health considered only as a cost and not an investment

e.g. cost savings from data driven care are not linked to the value of the data



Poor levels of public trust in the secondary use of health data

e.g. little understanding of innovations developed by industry through health data