Data Quality

Annie Jeffery
Population Health Intelligence Lead
NHS England, South

Kent Surrey Sussex
Academic Health Science
Network

BEAUTIFUL
INFORMATION

Why is it important?

Acceptable data quality is **crucial to operational and transactional processes** and to the **reliability of business analytics** / business intelligence reporting

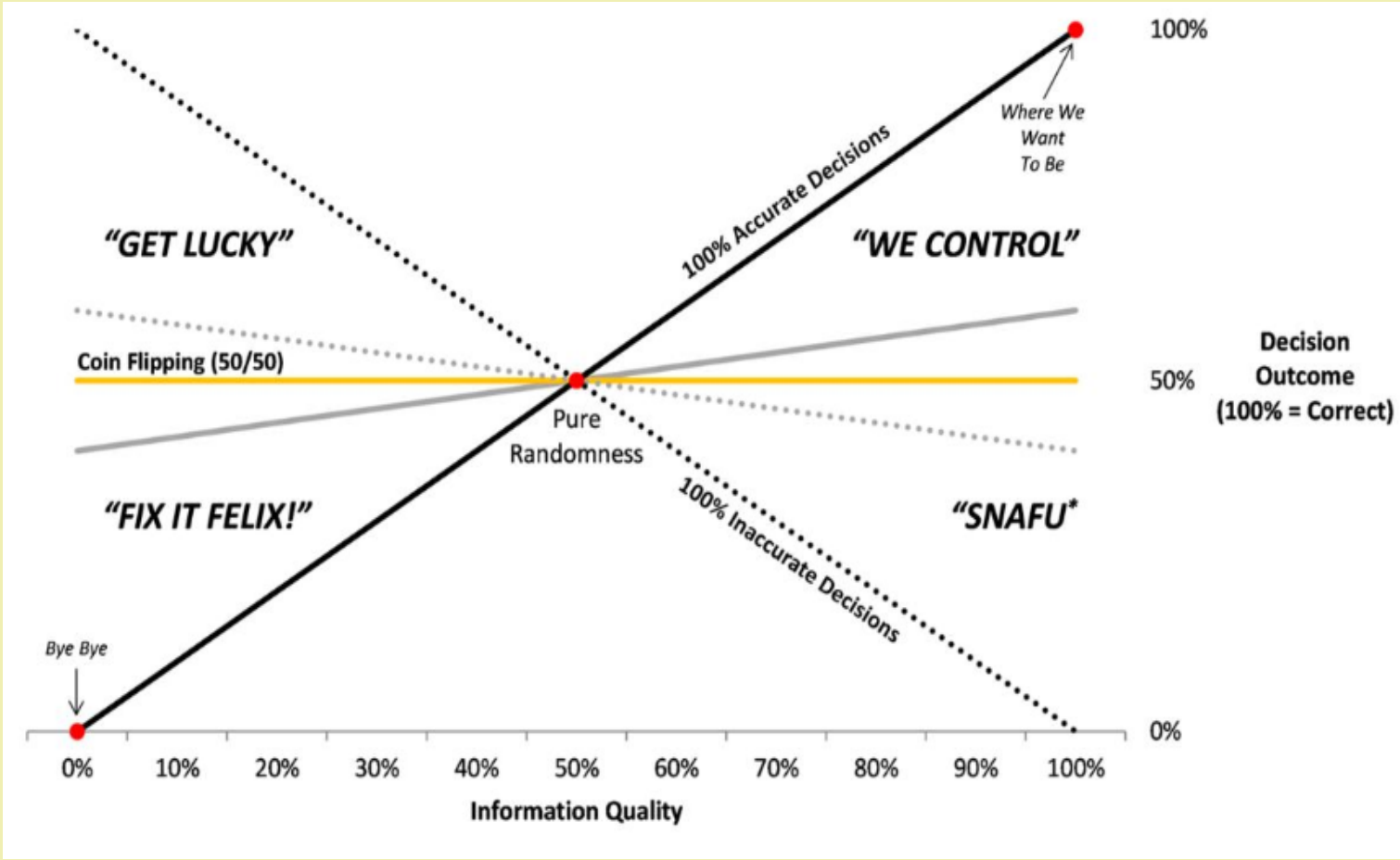
High quality information leads to **better decision making to improve patient care and patient safety**, and there are **potentially serious consequences if information is not correct and up to date**

Management information produced from patient data is **essential for the efficient running of the trust** and to **maximise utilisation of resources** for the benefit of patients and staff

Poor data quality puts organisations at significant **risk of: damaging stakeholder trust; weakening frontline service delivery; incurring financial loss; and poor value for money**

Mersey Care NHS Trust

Errors using inadequate data are much less than those using no data at all. Charles Babbage



Kent Surrey Sussex Academic Health Science Network

www.beautifulinformation.org

What is data quality?

Data are of high quality "if they are fit for their intended uses in operations, decision making and planning." J.M. Juran

Kent Surrey Sussex Academic Health Science Network

www.beautifulinformation.org

Scope of data quality

Data quality is not just about entry errors, and should be considered from design through to processing.

1. Data design

- Coverage
- Granularity
- Relevance

2. Data entry

- Accuracy
- Validity
- Default value %
- Completeness
- Standardisation

3. Data processing

- Linkage
- Consistency
- Timeliness
- Availability

Data design

Why are you collecting the data?

What will it be used to understand?

Is the data you are collecting fit for this purpose?

Coverage: Do you have data about every individual in the population?

Population – defined as a group of people – not necessarily population of England. Resident = everyone living in catchment area; registered = everyone registered with GP in catchment area; responsible = everyone under the care of service in the catchment area.

Relevance: Do the data that tell you what you need to know?

Could you consider proxy measures? I.e. prescriptions collected used for drug adherence. What quality issues does this introduce?

Granularity: Do the data provide enough detail to answer my question?

Where is person-level data necessary? What can we do with high-level or aggregate data?

Data design

How can we ensure the necessary coverage, relevance and granularity?

"I want to collect everything about everyone everywhere – that way I will be able to understand whatever I need to!"

The above statement is not possible and to try can be unnecessarily time consuming for data entry staff, data processors, analysts, decision makers…

A balance is needed between futureproof and practical!

Co-development with decision makers to understand a wide range of potential use cases, for now and in the future, will help set data requirements.

From data requirements data for collection can be specified. This will ensure the data collected has the best chance at having a suitable coverage, relevance and granularity.

User Requirement: Understanding the health needs of our population

ICS users will also need to understand the health ‘need’ of their population including, targeting where they plan interventions

User need

Example users: National Directors, Regional Directors, ICS Leads

Objective & Hypothesis to investigate

User acceptance / Key Success Criteria

What we need from stakeholders

How this could look in the dashboard

Work effort estimate

Work area	Level of challenge (1-5)	Comment
Data acquisition	2	NCDR data available to be included
Metrics development	3	Support required to agree metric definition
Visualisation development	3	70+ metrics on a page could get confusing
Approximate time	4 weeks	
Approximate cost		

A note about coverage…

How can we ensure we get the data about everyone we need?

1. Opt-outs

- By 2020 all health and care organisations are required **to apply national data opt-outs where confidential patient information is used for research and planning purposes**.
- "Planning purposes" = anything that is not direct care for an individual
- **Pseudonymised data is not exempt**
- Opt-out rates are subject to change according to public mood!

2. Out-of-area

- Students at university
- Students returned from university
- Relocated with work
- Attending specialist services
- Retired
- Holiday makers
- National and international migration

3. Out-of-contact

- Moved to the area and not re-registered with a GP
- Never been seen by health services
- Particularly relevant for groups who may have inequitable access to healthcare

Data accuracy

Are the data a true reflection of the objects they describe?

Data accuracy is difficult to measure; easier to estimate….

Ways to estimate data accuracy:

- **Data verification checks** on a sample of data
- **Validity** – do the data entered reflect **values that are possible** in the real world?
- **Defaults** – have default values, e.g. “unknown” been entered **more than would be expected?**

Data verification checks

Problem 1: We can't check everything:

- **Check against expected values** based on known trends
- **Identify commonalities** in unexpected data:
  - Are all unexpected values from the same variable? I.e. is the recorded height always a bit weird?
  - Are all the unexpected data coming from the same hospital?!
  - Logs of when data items are added/updated can show if the unexpected values are all entered by the same user or during the same time frame (e.g. night shifts
- **Target unexpected values with something in common** – this is more likely to be a true error with a non-random cause

Problem 2: We don't always have access to the true value:

- Transcription of data, e.g. from written medical notes, should no longer be taking place, as this has an increased chance of error in entering data; **data should be entered electronically in real-time**
- **Checks can be made against other sources** if the same data are entered in more than one source; however, a hierarchy for the correct version of the truth will need to be established
- Bear in mind, **real-world changes may have occurred since data entry** – data should reflect the truth at time of contact
- **Do patients have a role to play in verifying their own data?** How much credibility will the patient have as the correct version of the truth (they may misremember or misunderstand)

Data validity

Do data match values that are possible in the real world? I.e. all heights should be under 3m

Do data match values that are possible given other data ? I.e. men cannot be pregnant

Dataset design:

- Validation and business rules should be specified when the dataset is being designed, or any new data items added

Data cleaning option:

- Invalid data is identified during processing
- Staff responsible for data entry alerted and correct value corrected (if known!) – at source and in extract!
- Where unknown, data can be amended to the default value (e.g. unknown)

Validation rules option (catching errors in real time!):

- Data entry software can prevent users entering data that does not fit valid values
- Data entry software can prevent records from being saved, where invalid values exists
- Data entry software can highlight invalid values to users upon entry
- Data submission software can prevent data being submitted if invalid values exist
- Rules can be based on individual data fields or based on values in other data fields, e.g. if sex = man; pregnancy cannot = true

How can we influence suppliers to develop automated data validity rules and checks in data collection software?

Clinical involvement in data quality improvement

Research shows that that one of the reasons for poor data quality is a lack of clinical involvement in the collection, validation and use of the data

Data design:

- Is the dataset clinically relevant? Does it provide clinicians with information that is useful to them?

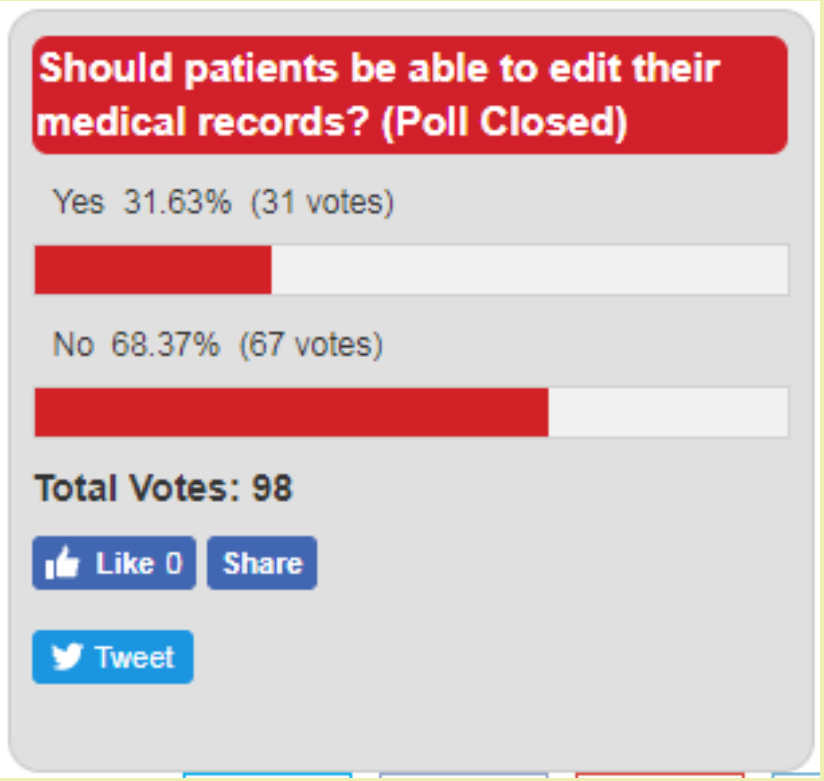
Clinical engagement in immediately relevant data:

- User-friendly, self-service dashboards, co-designed with clinical teams can encourage interest in the data
- If clinicians are not engaged in data accuracy checks from the beginning, they are likely to lose faith in the data (vicious cycle!)
- Support may be needed to kick off use of the data and involvement in verification

Clinical engagement in wider data quality:

- Moving to whole-system thinking, what is the wider impact of poor data quality?

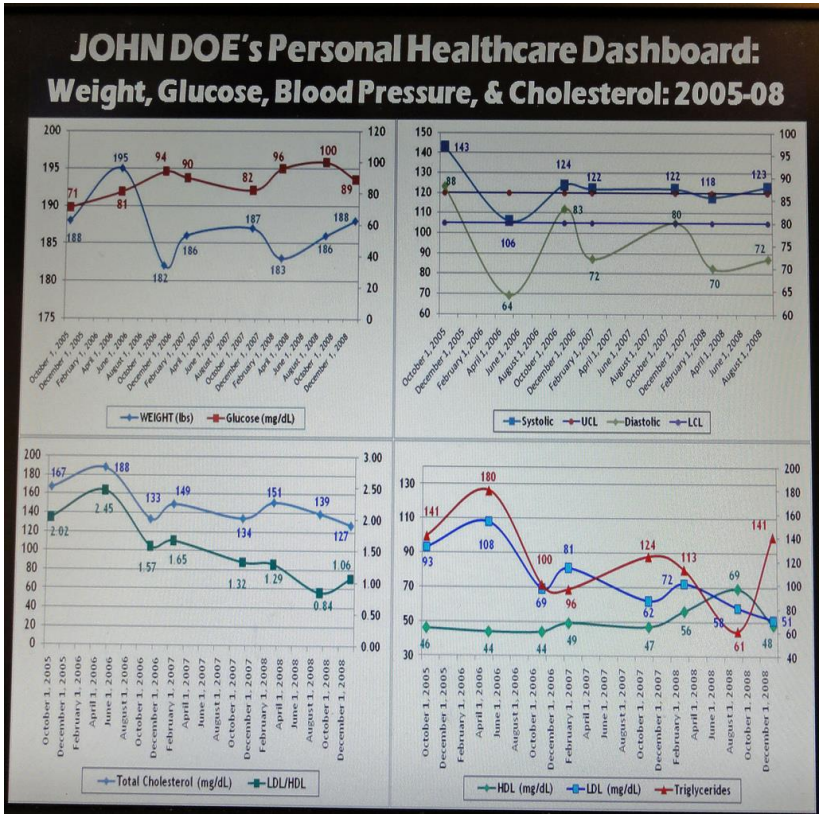
Can patient involvement improve data quality?



Source: GP Online

Patient involvement

Patients can currently access GP and Summary Care Records. They cannot edit their record, but they can suggest changes if they believe an error has been recorded.



Patient dashboards can engage patients in their individual health data – this can encourage patient checks of data accuracy.

Capacity to investigate errors and update patient records is needed.

How will patient-spotted errors be verified?

Psychological impact of health and treatment data presented without a professional explanation of context must be carefully considered.

What are national teams doing to help?!

Data standards

National datasets:

- NHS Data Dictionary (description, units, change log)

Local datasets:

- Should have clearly specified and accessible definitions
- Co-development of a data dictionary with those who enter and use the data is beneficial
- Definitions should be reviewed to ensure they continue to accurately describe the variable and possible values
- Units or measures should be clearly specified
- Changes to the data collection form, definition or possible values should be logged

All data entry staff should be familiar with data dictionaries and definitions

Data Quality Maturity Index (DQMI)

The DQMI methodology is in its 6th iteration and has been stable for five successive quarters. It measures the following DQ dimensions (see Annex A for descriptions) across eight datasets:

- Coverage
- Completeness
- Validity
- Defaults

Though not an official publication the interactive DQMI reports are regularly used by providers to support their own DQ work.

(average 27 visitors / 104 page views per day following DQMI May 18 publication)

The table below shows the relative DQMI performance over the past five reported quarters.

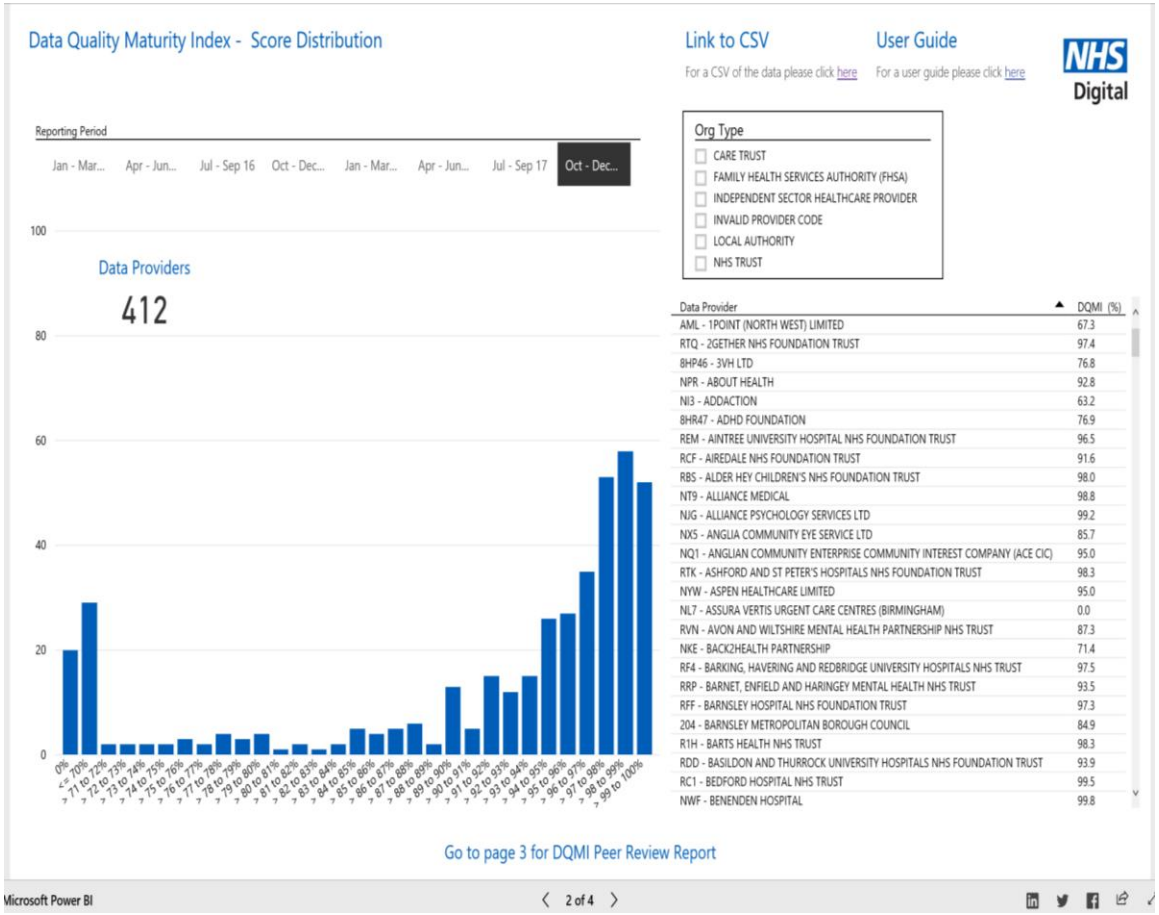
Period	AE	APC	OP	DIDS	MHSDS	MSDS	IAPT	CSDS
2016-17 Q4	99.0%	96.3%	97.4%	95.1%	95.9%	95.0%	96.6%	93.5%
2017-18 Q1	98.9%	96.7%	97.4%	94.9%	96.0%	95.2%	97.3%	93.5%
2017-18 Q2	98.9%	96.3%	97.0%	95.2%	95.0%	95.8%	97.7%	93.5%
2017-18 Q3	98.6%	96.4%	97.3%	95.1%	95.1%	95.5%	97.7%	93.9%
2017-18 Q4	98.6%	96.3%	97.1%	95.2%	88.8%	95.5%	97.6%	94.6%

***AE**=Accident & Emergency **APC**=Admitted Patient Care **OP**=Outpatients **DIDS**=Diagnostic Imaging Dataset **MHSDS**=Mental Health Services **MSDS**=Maternity Services Dataset **IAPT**= Improving Access to Psychological Therapies **CSDS**=Community Services Dataset*

Data Quality Maturity Index (DQMI)

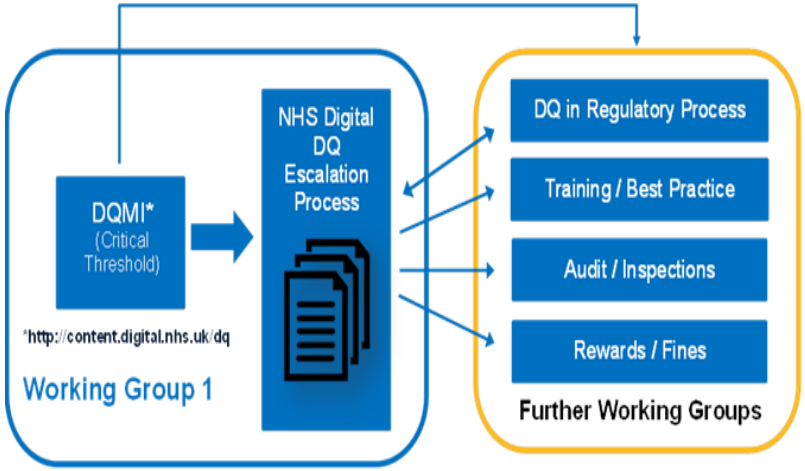
Each quarterly DQMI publication includes the DQMI Methodology) which in turn is supported by the provision of both the underlying data (as a CSV) and an interactive report using PowerBI

Future enhancements to the PowerBI report are planned including the provision of time series data.



NHS Digital Incentives and Levers Programme

- The Incentives & Levers Programme is a **joint collaboration** with NHS England, NHS Improvement, CQC and Public Health England.
- The main output to date has been the DQMI and the **Escalation Process** which enables data collections teams and the ALBs to use NHS Digital's Contact Centre to raise and **track DQ issues** with providers.
- **Examples** - The Escalation Process is currently being used to
  - Address coverage issues relating to 26 Mental Health Services providers (representing 3% of total activity)
  - The process has also been used to support the NHS Chief Nursing Officer to address DQ issues with Maternity Services data;
  - To assist NHS Improvement Estates and Facilities with the 2017/18 ERIC collection
  - In coordination with the Acute EPR Red Team.



Escalation Activity	DQ Contact Letter
Letters Issued in Last Period (2017/18 Q4)	19
Responses Received in Last Period (2017/18 Q4)	16
Letters Issued in Period (2018/19 Q1)	43
Responses Received in Period (2018/19 Q1)	24
Providers Not Responded	22
Planned Escalations in Next Period (2018/19 Q2)	22

Revised NHS Digital DQ Strategy

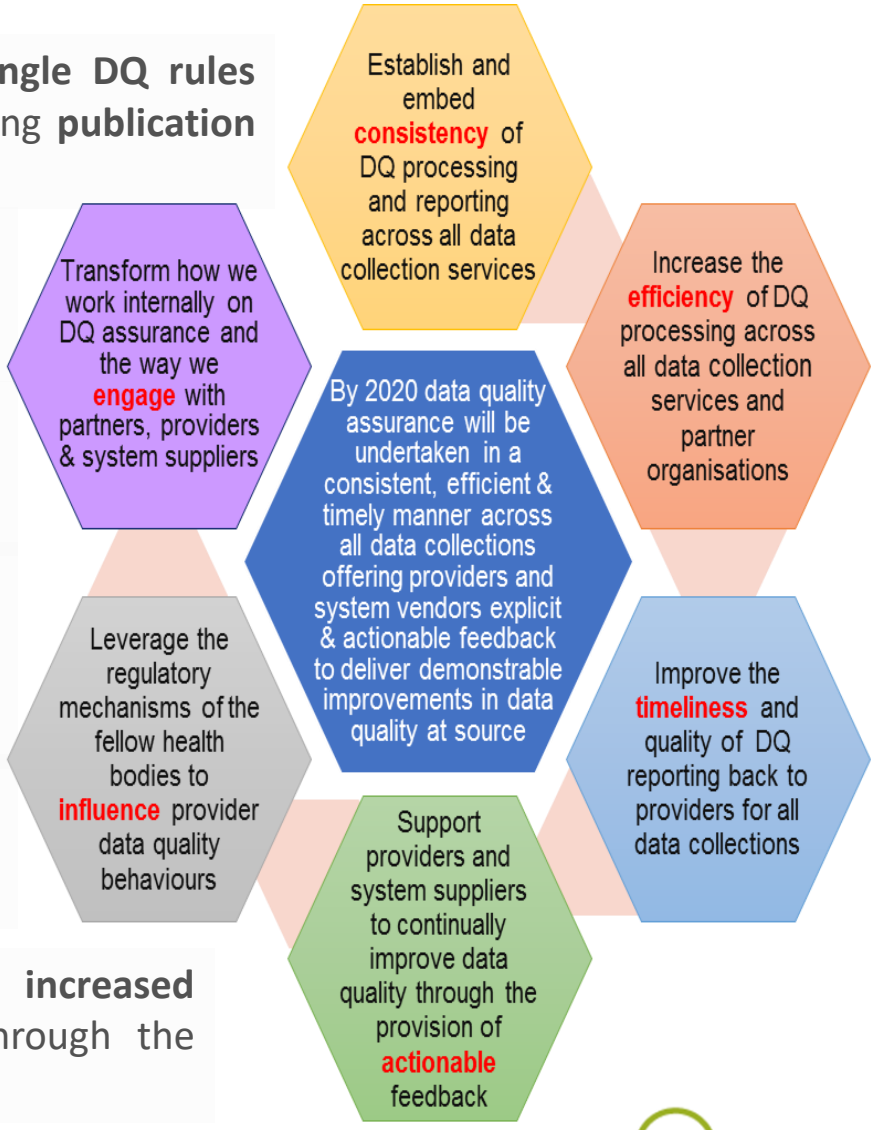
Consistency central to achieving strategy – using a **single DQ rules repository** within Data Services Platform (DSP) – enabling **publication of rules** to providers and vendors

Efficiencies realised through DQ processing at **point of submission -** removing subsequent processing by health partners and customers

Timeliness of DQ reporting essential to **improving DQ at source** – surfacing issues in **near real-time** supporting immediate action by providers

Actionable feedback in a standard, **clear and concise format** required to **focus providers** on critical DQ issues at source – **Data Access Environment** key component in this.

Influence through **contractual and regulatory incentives and levers** will ensure providers understand their **obligations and the consequences** of not taking action.



Engagement with partners, providers and vendors **increased** through **refocusing NHS Digital resource** released through the technology capabilities above

What about data linkage?

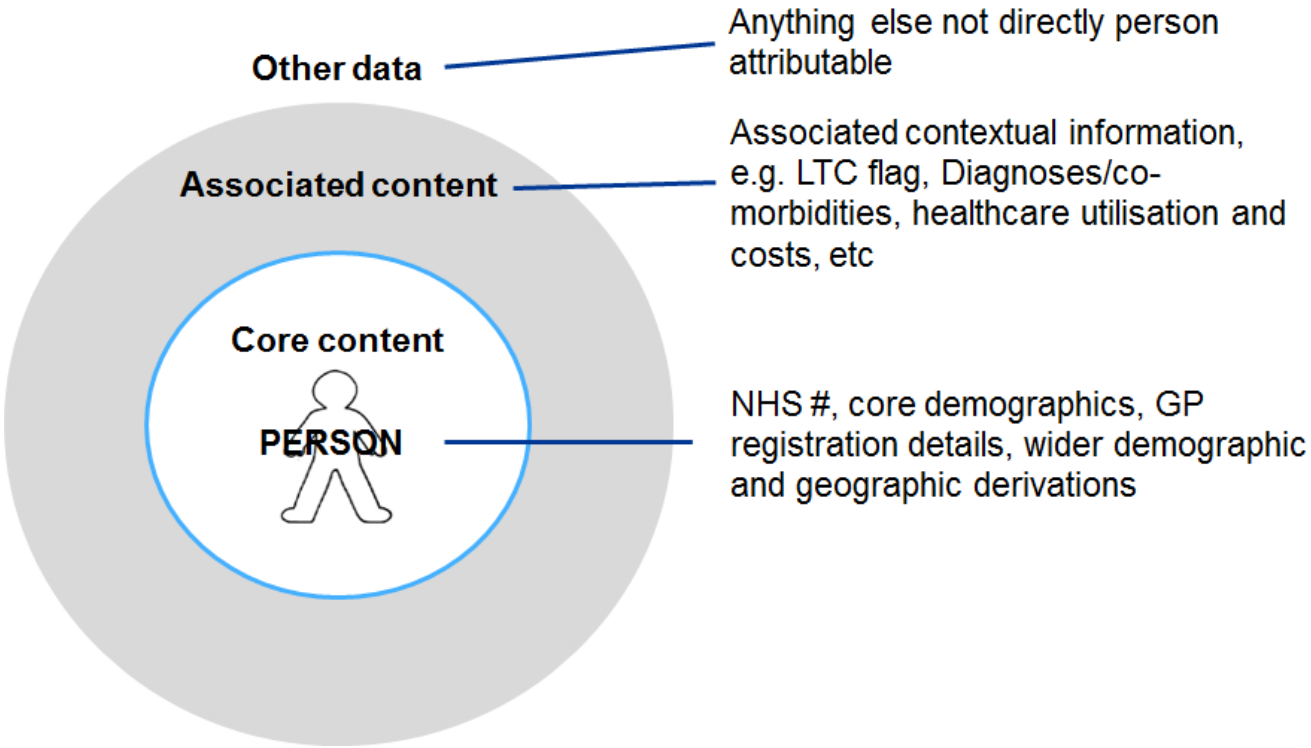
Master Person Index

A master person index is data and information used to maintain consistent, accurate and current demographic, and essential clinical, data on persons.

They are intended to:

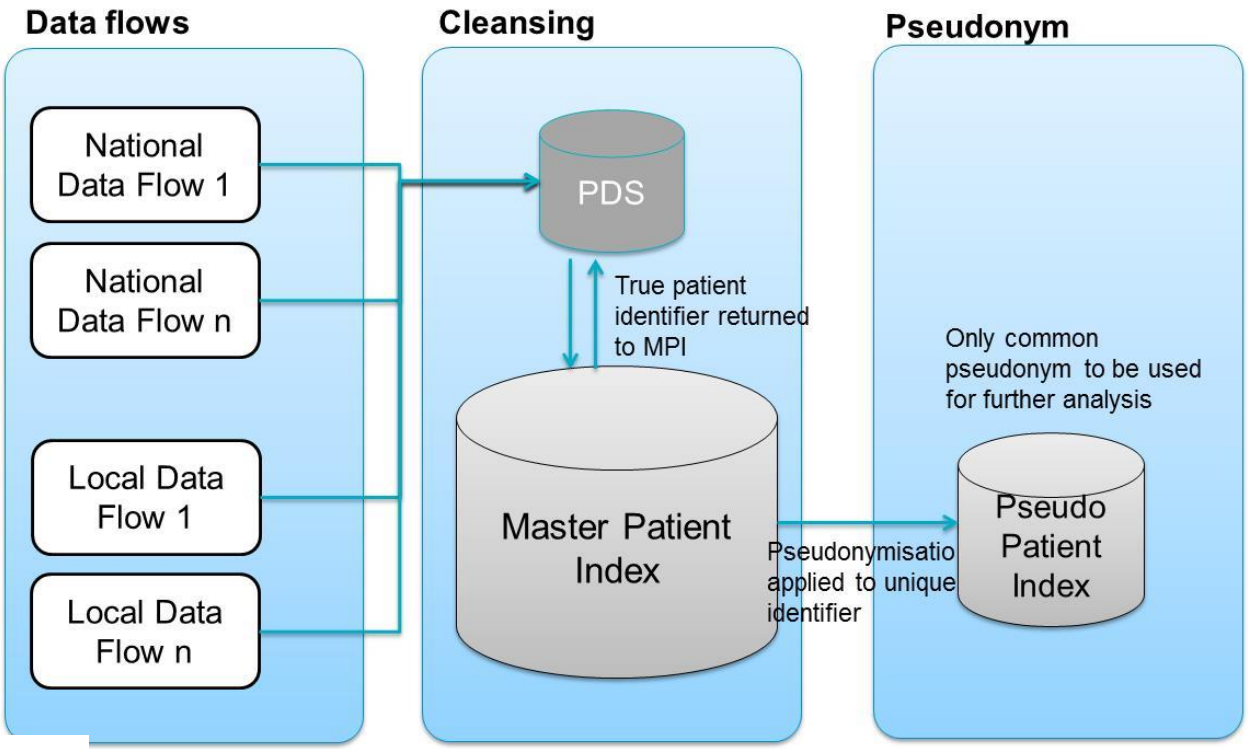
- Solve the common problem of multiple systems and datasets gradually becoming inconsistent with respect to the person's most current data
- Enable improved data matching and subsequent data linkage across datasets

More detailed content in supporting Annex



NHS England are working to…

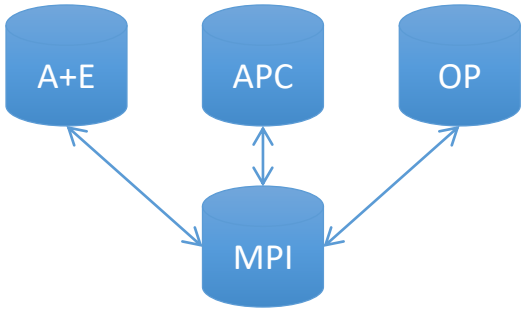
- Improve the identification of persons to facilitate linage of data by:
  - Tracing data against a reliable patient list (e.g. PDS, NHAIS)
  - Leverage details not in patient lists to identify further matches (e.g. Local ID)
- Build on opportunities to improve person matching
  - Applying weights/scores to deterministic steps based on probabilistic principles



Better linkage and quality of linkage

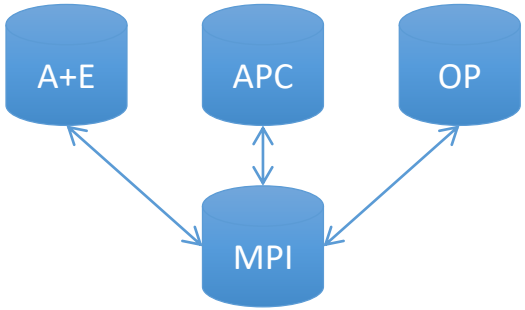
a) Across datasets using NHS #

c90% records



b) Through MPI using NHS #

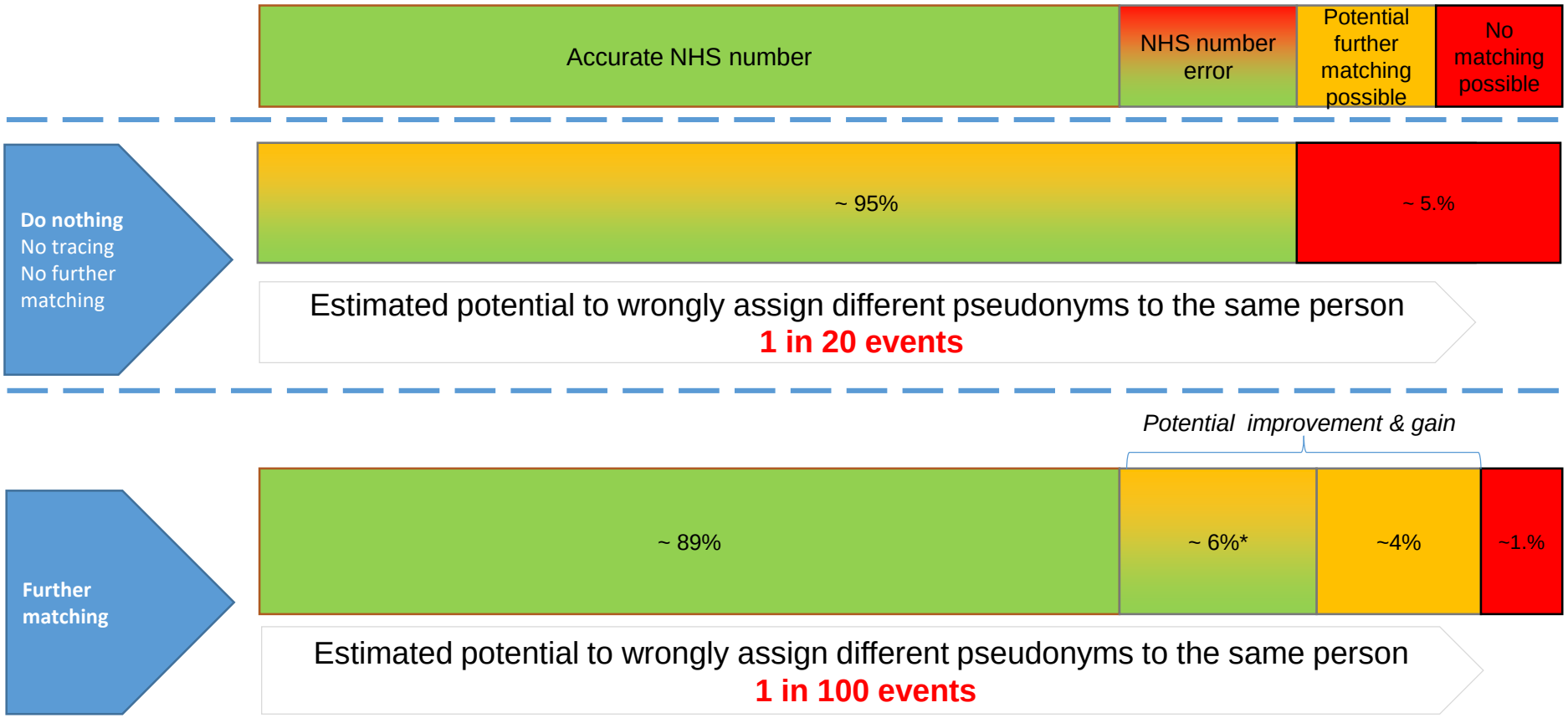
c95% records



c) Through MPI using algorithm xx

c97% records

Further matching

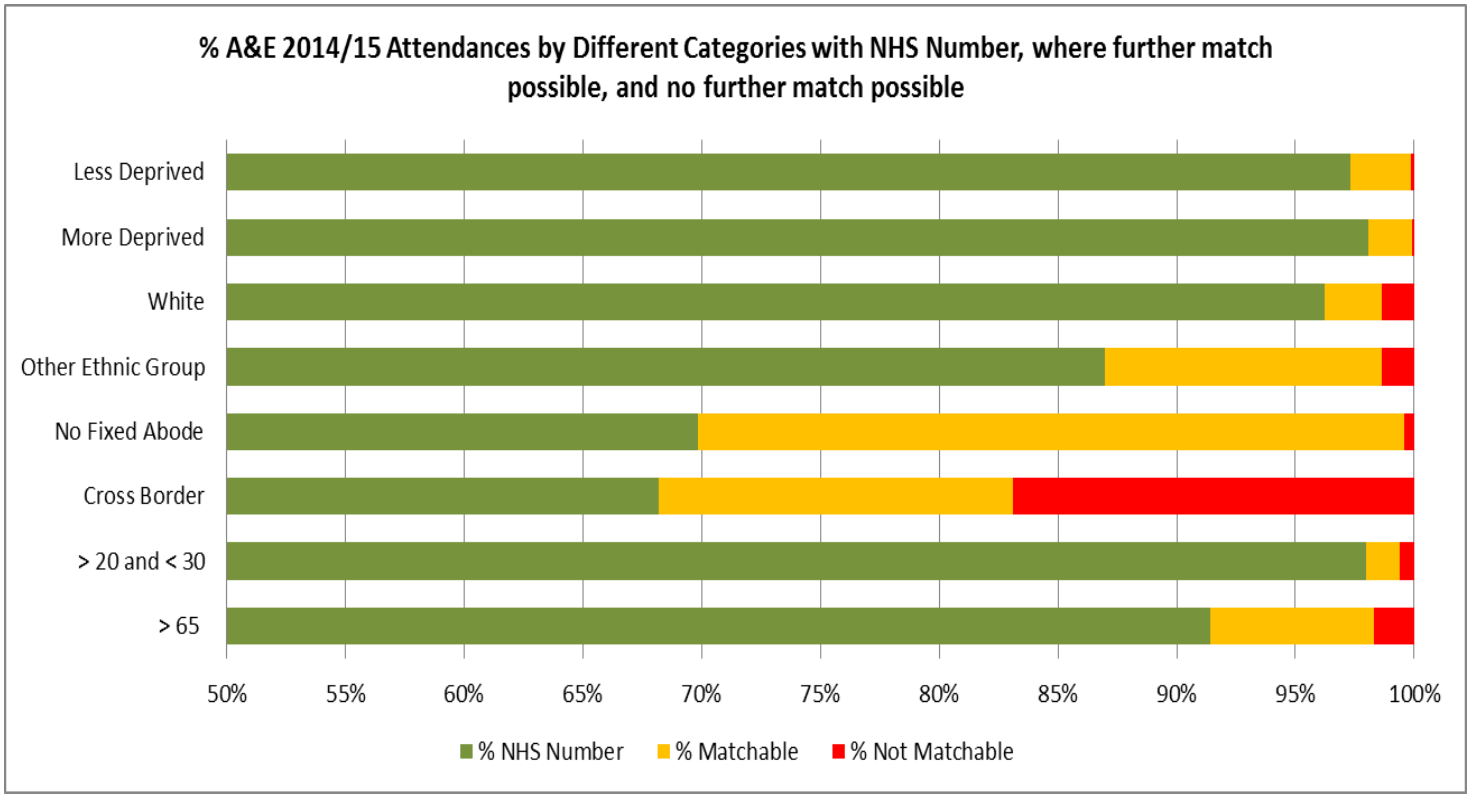


- Estimates based on established national datasets
- Further matching possibilities based on combinations of Local ID and provider code, Date of Birth, Sex, Postcode, GP practice code

Kent Surrey Sussex Academic Health Science Network

www.beautifulinformation.org

Health Inequalities and NHS Number Matching



Kent Surrey Sussex Academic Health Science Network

www.beautifulinformation.org

Questions?