NHS Connecting for Health	Research Capability Programme (Enabling Phase)			
	Programme	*NPfIT*	DOCUMENT RECORD KEY	
	Sub-Prog / Project	*Research Capability Programme*	*NPFIT-RCP-DS-0002*	
	Prog. Director	*Peter Knight*		
	Owner	*RCP Programme Board*	Version	*PV*
	Author	*A Coulton*		
	Version Date	*15th August 2008*	Status	*Publication*

Research Capability Programme

Data Quality, Standards and Linkage

(PD12)

Supporting paper on Data Quality Metrics and, Potential Measures and Dashboards

Amendment History:

Version	Date	Amendment History
0.1	11/04/08	Initial Document Structure
0.2	12/04/08	Revised content
0.3	29/16/08	Pre-publication - amended for standard layout
PV	15/08/08	Publication version

Reviewers:

This document has been issued for review as follows:

Name	Title	Date of Issue	Version	Notes

Approvals:

This document requires the following approvals:

Name	Signature	Title	Date of Issue	Version

Document Location:

This document is only valid on the day it was printed. Please contact the Document Controller for location details or printing problems.

This is a controlled document.

On receipt of a new version, please destroy all previous versions (unless a specified earlier version is in use throughout the project).

Related Documents:

These documents will provide additional information.

Ref no	Doc Reference Number	Title	Version
1	NPFIT-NPO-GEN-IP-0067	Glossary of Terms Consolidated.doc	Latest
2		Government response to the health committee report on the electronic patient record	Nov 07
3		UKCRC report	June 07
4	NPfIT-RCP-PID-0001	Research Capability Programme – Enabling Phase, Programme Initiation Document (PID) -).	
5		Best Research for Best Health	
6	NPFIT-FNT-TO-EAR-0031	NHS CFH Introduction to the Enterprise Architecture	
7	NPFIT-NCR-SUSPBR-0044.02	NHS CFH Data Warehouse Reference Architecture	2.00

Contents

1. Purpose 4
1.1. Introduction 4
1.2. The Data Quality, Standards and Linkage Work Stream 4
2. Metadata, Data Dimensions and Data Quality 5
2.1. Introduction 5
2.2. Metadata about Data Quality and Metrics 5
2.2.1. Introduction 5
2.2.2. Suggested Data Quality Metrics 5
3. Presenting Information about Data Quality 8
3.1. Accessing Information about Data 8
3.2. Finding Information about Data Sources 8
3.2.1. Index 8
3.2.2. Search Functions 8
3.3. Dashboard Functions 9
Appendix 1 – Data Quality Metrics and Potential Measures 11
Appendix 2 – Dashboard Functionality 17
3.3.1. Xcelsius Dashboard 18
3.3.2. SUS 'Data Quality Report' (DQR) 20
3.3.3. eDQRS on SUS 21

1. Purpose

1.1. Introduction

This document is part of a series of 2021 products (PD00 to PD20) that make up the Enabling Phase of the Research Capability Programme (RCP). The Programme comprises six work streams:

- Technical Architecture
- Functional Scope and Business Case
- Data Quality, Standards and Linkage
- Information Governance and Threat Assessment
- Infrastructure
- Communications and Stakeholder Engagement.

These work streams are highly inter-related and the products of the individual work streams should not be read in isolation. For example, many of the products will highlight Information Governance issues which may not be directly addressed by the specific product itself, but which will be addressed by the relevant product from the within Information Governance and Threat Assessment work stream.

Further information on the overall RCP providing the background and context for this document is included in PD00.

1.2. The Data Quality, Standards and Linkage Work Stream

The purpose of the product PD12 was to flesh out the processes and components in PD11 Approach to Data Quality, Standards and Linkage in order to:

- Propose detailed examples of required data quality measures, standards and linkage services required to support research; and
- Provide input into the Business Case (PD10).

Work stream 3 (Data Quality, Standards and Linkage) aims to identify the infrastructure and mechanisms required to ensure that the data within the sources that will be included in the HRSS are 'fit for purpose' for research needs.
The additional complementary products to PD12 are:

- PD11: 'Approach to Data Quality, Standards and Linkage' Report: document. Documents the approaches to be followed to ensure that NHS data areas of an appropriate quality for research purposes, building upon the data quality, standards and linkage work that is already underway for other purposes.
- PD13: 'Threat Assessment Report'. The purpose of this product is to examine the potential threats arising in the process of data linkage.

PD12 therefore builds on and elaborated the PD11 Approach to Data Quality, Standards and Linkage. It documents the work that must be undertaken on data quality, data standards and data linkage to ensure that data are usable for research purposes.
The document included includes Sections on the metrics and metadata needs for assessing data sources, quality and standards, and ways of presenting information about data quality. However, more material was generated on data quality and associated presentation methods than was required to meet the initial remit of PD12. The content is still of considerable value and relevance to the future development of the Research Capability Programme and developing the HRSS. Therefore this supporting paper has been developed. It includes the Sections from PD12 that relate to Data Quality and Presentation, namely Sections 3 (but omitting Sections 3.2 and 3.3) and 4, and then presents the additional information as Appendices.

2. Metadata, Data Dimensions and Data Quality

2.1. Introduction

Researchers need to be able to judge the potential value of the data sources provided through the HRSS. In order to do this, they will require information about the properties of that data ("metadata"). Examples of properties that can be used to judge data value include:

- Source population basis (e.g. geographic coverage)
- Person- based (whether data is person- based or aggregated)
- Record identification (whether the source data is identified, pseudonymised or anonymised)
- Linkable (whether records are linkable, e.g. use of the NHS Number)
- Data standards (conformance with known data standards)
- Data quality rating (data quality procedures and measures in place).

These data dimensions are explored in more detail in this Section.

2.2. Metadata about Data Quality and Metrics

2.2.1. Introduction

The HRSS should provide researchers with a series of DQ metrics developed specifically to support research. Researchers need to be able to compare the quality of data sets and data (either from established data flows or data specifically commissioned for a research programme) using a consistent set of data quality criteria. This will enable them to judge whether or not the data would suit their needs, in particular whether the data are of sufficiently high quality to be used within their research programme. In order to meet the business requirements of clinical research, DQ metrics must comprise not only those attributes of data quality which are task-independent i.e. that reflect the status of the data without the contextual knowledge of the application[1] (e.g. completeness, validity etc), but also those applicable to the particular research topic and context.

Researchers would request and obtain information about the data via the HRSS portal or API. This would act as the interface between the researcher and their data that they may wish to use. The information about the quality of data could be obtained by a search for the data quality ratings and other quality attributes about the data, or using the dashboard for display of values for the DQ measures of the datasets or data requested, based on the DQ metrics.

As well as aiming for consistency in design and presentation of the metrics, there would be provision within the HRSS for their periodic reassessment (either regularly or by some kind of manual audit, that may involve IG permissions). It is intended that researchers would be directly involved in all these processes. The measures and findings would also have to be fed back to the providers of the data to fuel data improvement.

2.2.2. Suggested Data Quality Metrics

There is a wide range of possible metrics for describing data quality and for use as metadata. Some proposed data quality metrics for use within the HRSS are listed in Table 1. These are presented as a non-prioritised list. Issues that would require further consideration for the actual measures are documented in Appendix 1 – Data Quality Metrics and Potential Measures. The feasibility of providing these for different data sets will be assessed in the ongoing work of the Programme.

[1] Pipino, Yang W. Lee, and Richard Y. Wang: Data Quality Assessment. *Communications of the ACM 45, (April 2002).*

Table 1. Metadata for Data Quality

Data Quality Metrics	Definition (HRSS)
***Data attributes*:**	
• Data components	List of classified subunits of information, also known as 'fields', specified for use (in a research project),
• Completeness	All data fields are populated with data of sufficient breadth and depth for the task in hand ; or– Or - All data fields are populated with data values sufficient for use for their intended purpose.
• Validity	The data conform to the defined format, structure and defined values for the fields designed to contain them when collected.
• Accuracy – not yet included in initial stages of the HRSS	The data represents the actual reality.
• Coverage	All relevant categories are included.
• Inclusion of identifiers	The extent to which the data include items which identify the individual.
• Treatment of identifiers	The arrangements operating for processing of and access to identifiers.
• Data derivations	(Suggested definition for the HRSS context): The extraction of data items from sources / other items or forms of processing, in a manner which produces an output of defined integrity without compromising the entity or disclosing the identity of the source.
• Indications of data derivation	Indications of derivation processes applied - single or multiple – and where multiple routines existed, which stage in the derivation process had produced the useable output.
• Possibility for duplicate, confused or merged records	Indications of the presence of identical copies of records in addition to their original. Records in which components have become disconnected and attributed randomly across more than one patient, resulting in a loss of the record integrity, and where data from more than one record have been combined.
• Default	Data item or code used to provide a component with a valid structure and format as a substitute where the data item is not known or not available.
• "Consent flags"	An indicator that the patient has given their consent to use of their identifiable information for the purposes of the research project.
***Coverage attributes*:**	
• Geographic coverage	A measure of the precision of geographic coordinates used and/or a measure of the geographic coverage of the data set.
• Age range	The difference between the required minimum and maximum age limits of the cases
• Sample rate	The measurement of the frequency with which snapshots at a given point in time are taken over a continuous progression (of time).
• Sample type	Defines the format of a single sampling value (used to test hypothesis by taking a small part of something intended to be representative of the whole)), as a – and measure of usefulness with reference to the research.
• Gender	Classification of objects approximately corresponding to the two sexes and/or other gender identifiers.
• Condition-specific	State of health applicable to the subject of research.
• Time-related	Considerations that relate to how up-to-date the data are with respect to the task for which they need to be used, and time properties inherent

Data Quality Metrics	Definition (HRSS)
	within the data content.
• Timeliness	Speed of dissemination of the data - i.e., the lapse of time between the end of a reference period (or a reference date) and dissemination of the data.
• Temporal coverage - (starting and ending dates, length of history)	Identification of a span of time for data collection and its end-points.
• Frequency of updates	Measurement of the number of occurrences within a given time period when more recent data / information is supplied.
• Last update	The most recent replacement (of data) to reflect any changes since the previous version.
***Coding**:* (including clinical coding)	Coding is the conversion of information on a subject or unit into another form according to a defined set of rules, often for the purpose of data storage. Clinical coding is the assignment of clinical activity in different settings in accordance with a disease and procedure classification system, which permits the systematic recording, analysis, interpretation and comparison of morbidity and procedural data.
• Coding scheme	The parameters assigned to the coded values used in the design of specific coding schemes / rules.
• Coding validity	The data conform to the defined coding structure.
• Coding completeness	The data have been recorded in all the required coded fields to at least the minimum required specificity.
• Coding depth	Use of the range of available digits in the classification code structure to present degrees of clinical complexity and dimension.
• Specificity of coding	The application of a clinical code at the most atomic level for maximum clarity.
• Number of diagnoses/ codes per entry	Presentation of multiple clinical details.
• Audit of coding	Independent review of coding to assess the application of the classification rules and system controls. Examination and verification of coded data with reference to professional standards, where these exist.
Information about the data quality assessment itself	Summary details about the DQ assessment carried out supported at the portal.
When completed / updated	Details would be provided on: dates for completion and update of the assessment; identity of person(s) who completed the assessment; the assessment method used; and the review date.
By whom completed / how	
Date for review	

3. Presenting Information about Data Quality

3.1. Accessing Information about Data

The HRSS service would need to support users with a range of needs and at different stages of their work, from those who presented specific requirements to others who might only be able to refer imprecisely to an area of interest. Clear presentation of well-defined metadata would also help researchers specify the descriptors that could assist with locating and identifying the information they needed.

This Section describes the functions that the RCP is proposing to be provided to enable researchers to identify the suitability and availability of data that they hope to use in their research. The researchers could then apply for and obtain information about the quality attributes of the data itself and reference the relevant metadata. This would provide them with the interface with the information they would need to determine the suitability of the data for use in their research, subject to the requisite governance controls.

It is proposed that these functions would be available as a catalogue of search facilities and via a dashboard presentation on information about data sources and all the other data dimensions identified in PD11, Section 6, and the data itself. Both search and dashboard functions would need to provide information in real time and on the frequency of updates, as applicable.
For simplicity, this Section presents the functions as a web portal, although additionally programmatic access to the functions may be possible in addition e.g. via an API (although this is not described in detail in this document).

Where dashboard or programmatic access are not sufficient, researchers should be granted direct access to individual experts with the requisite knowledge to assist with the data content and aspects relevant to its use in research, including data quality, standards and access permissions. This type of service is available, for example, for users of the General Practice Research Database (GPRD) where users can contact members of the research team for guidance on complex questions relating to the applicability of the data to their research needs.

3.2. Finding Information about Data Sources

3.2.1. Index

An index of all portal services, providing a summary list with a brief description of search and dashboard properties, would be available at the point of entry. This would show the level of access to identifiable information associated with each user's research permissions, triggered by their access pass codes at the point of entry.

3.2.2. Search Functions

Search facilities applying the most advanced forms of search engines would support researchers in calling up the information they needed. This would enable the researcher to search, for example, for the items described below:

- Person records, data or data sets, or combinations of these, to detailed individual item-level (subject to permissions). The researcher could also specify their research criteria, including human subjects, to call up options that could support their requirements. This search facility would have the potential to access all records, including data and datasets currently in use for clinical research, and combinations of these, or data not previously used for research, . It should access data sources routinely available, as well as new or bespoke sources.

- Other web-enabled directories of databases, such as "DoCDat" held by the Information Centre, including the lists of the databases they hold and information they provide about the databases, and complementary data including their assessment of their quality and suitability for use.

- Information about the value of the data sources for research, where available, for example: published papers by previous researchers in the field and feedback from other researchers.

- Information on the data itself and any relevant contextual information from world-wide reference libraries.

- Information on all known and available data sources held nationally and locally, from data sources routinely available, to new or bespoke sources that may not have been identified previously for research use. Metadata on the data sources which indicated properties, such as whether they were electronic, person-based, provider organisations etc would also be accessible.

- Examples of some of the better known data sources include:
- (Human subjects): the UK Biobank epidemiology study ('Report of the Research Simulations' by the UKCRC Advisory Group for Connecting for Health', Annex A, Section 3).
- (National data collections): Clinical Audit Datasets (e.g.: cancer, heart disease, renal disease and diabetes supported by the National Clinical Audit Support Programme (NCASP) at the Information Centre; Hospital Episode Statistics (HES),); and Commissioning Datasets ion the Secondary Uses Service (SUS).
- (Primary care clinical records): GPRD and QResearch.
- Local sources could include patient records held on local Care Record Systems implemented in hospitals as part of the National Programme for IT (NPfIT) and patient records in General Practice systems.)

- Standards, wherever these exist for the required data, or more generally for data and information across the UK. Where these include standards published in Data Dictionaries for the NHS in England, Scotland and Wales, Scotland and N. Ireland this would also need to identify divergences in standards between UK devolved administrations; e.g.: the code for ethnicity has a different name and list of values in each of the three Data Dictionaries. Metadata would need to reference existing standards or indicate where they were loosely applied, (as in free-text fields), or do not exist.

  There would also be a need to specify for the HRSS all areas where standards do not currently exist but are needed to for the system to function in accordance with RCP requirements.

- Data Quality metrics in use in the HRSS, e.g.: completeness, validity, coverage, default and the ratings following the application of the metrics on the subject matter of the search where applicable. Metadata providing the information on the description and properties of the metrics are listed in Appendix 1 – Data Quality Metrics and Potential Measures.

  Facilities to display the information on rated values would be provided additionally in the dashboard – see Section 3.

- Information on Data Linkage, with a description of the linkage, identifying which items had been linked, the quality of the assessment and the impact upon data quality.

- Conditional, retrospective and prospective searches on aspects such as changes over time, and impact on content and quality attributes.

The search facility would also need to support details of off-line storage, security arrangements and information governance permissions pertaining to the data, and well as the arrangements for providing researchers with the results of their searches.

3.3. Dashboard Functions

It is intended that the dashboard tool would have web-enabled functionality with real time access to information about data for research.

The dashboard would be used for graphic display of any information on research subjects, data, sources, standards, metadata and linkage, supplementing and complementing the search functions of the HRSS. It would have the facility to present information and generate reports in a wide variety of formats. These might need to be in the form of detailed descriptions rather than trying to use devices

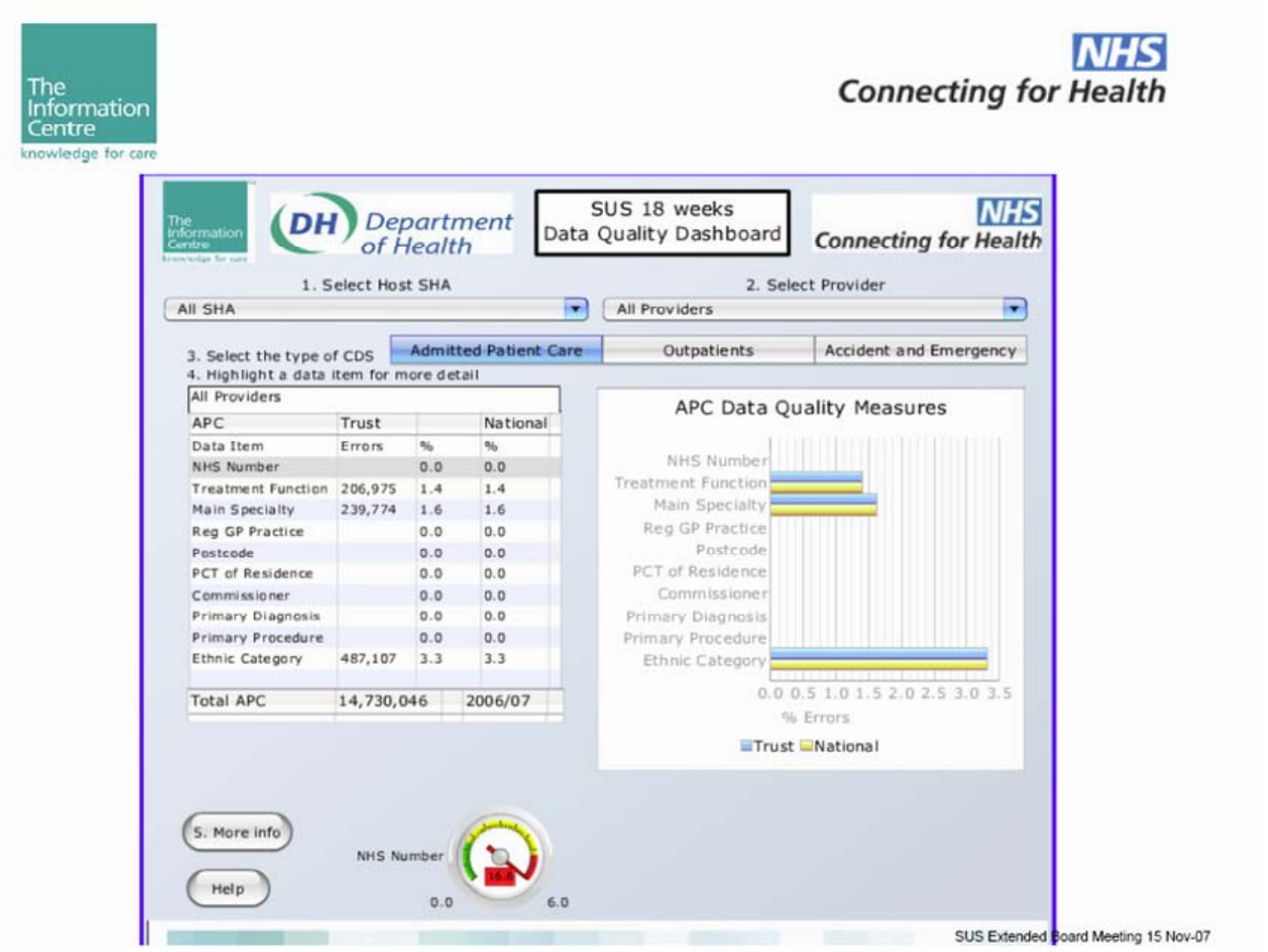
such as, for example e.g., “traffic lights”, especially where there are no consistent, objective criteria that can meaningfully be used to set thresholds. However, the graphic functions of the dashboard should be designed to accommodate a range of presentation options.

Its full potential would be realised in the display of reports giving values for the DQ metrics applied to the required items according to the user’s selection, or provided in providing listings of ready-made reports, for example, on some of the frequently used datasets such as HES or QResearch. These would support a range of requirements, from high level views such as quality and reliability summaries and trends of interest to the researcher, to the results of drilling down to detailed history and trend analyses for single items. The specification for the type of dashboard product proposed for the HRSS portal, with access to the range of information sources and configuration potential required for research, would involve significant development, considering what is currently available within the NHS. It would also need to support the defined user authorisation and permission levels, required for good governance. However, examples of similar applications already exist in some sectors, such as the ‘Cardiac Safety Dashboard’. This can track progress of studies based on updated data for participants in cardiology clinical trials. This is provided via the ‘Digital ECG Community’ portal (available from eResearch Technology)), which also provides eHealth Education distance learning capabilities.

The RCP would engage researchers in specifying the data, standards, metrics, metadata, linkage and other content, as well as the display features for the dashboard. The aim would be to provide a dynamic and interactive product capable of meeting both the most complex and the most basic needs of researchers, using the most advanced, powerful dashboard technologies.

A simple example of the use of dashboard techniques in relation to data quality is shown in Figure 1.

Figure 1. - Example Dashboard



Appendix 1 – Data Quality Metrics and Potential Measures

Table 2.

Data Quality Metrics	Definition (HRSS)	Considerations for DQ measures
Data attributes:		
• Data components	List of classified subunits of information, also known as 'fields', specified for use (in a research project),	The HRSS would need to reference the sources of all known metadata that exist for data, e.g.: national data standards published in NHS Data Dictionaries. Metadata about data would include the properties of the data in question, including where standards did not yet exist but which were needed for the HRSS to function as specified.
• Completeness	All data fields are populated with data of sufficient breadth and depth for the task in hand. Or: All data fields are populated with data values sufficient for use for their intended purpose.	Completeness as a data quality metric is generally used to define the extent to which data which meet their defined business criteria for their use are provided in all the fields specified, for example, in nationally published datasets. Criteria applied would include the presence of omissions, defaults, or non-valid coding.
• Validity	The data conform to the defined format, structure and defined values for the fields designed to contain them when collected.	Supporting criteria for validity would measure performance against defined standards for format and values, where those existed. The metadata would reference those standards, e.g.: published in NHS Data Dictionaries.
• Accuracy	The data represents the actual.	This is the 'gold standard' in DQ measures, but computerised systems have not yet convincingly demonstrated their suitability as a vehicle for supporting this. It is therefore expected that this would continue to be measured by audits and user feedback at least in the initial phases of the HRSS.
• Coverage	All relevant categories are included.	This would measure whether all categories which formed components of the subject area, (e.g.: conditions, specialties, or types of organisations), were represented in the selection process, and relative to what is expected, for example,: if the data source is all NHS- Trust types, NHS Foundation Trusts should be included.
• Inclusion of identifiers	The extent to which the data include items which identify the individual.	This would operate in the context that the use of person- identifiable data must be restricted to a justifiable minimum, preferably with the person's consent. Where there is a legitimate requirement for identifiers this would be expected to specify required identifier type: anonymised, pseudonymised, or person- identifiable. References would be provided to metadata supporting these descriptions. A hierarchy might be drawn up of identifiers for use in this context, in descending order, starting with the NHS Number. Where this was not available (as in Scotland and Northern Ireland) other demographic fields for use might include patient identifier, date of birth, address and postcode, person surname and forename.
• Treatment of identifiers	The arrangements operating for processing of and access to identifiers.	Controls would be applied to the access to identifiable information in accordance with approved governance arrangements. The metrics

Data Quality Metrics	Definition (HRSS)	Considerations for DQ measures
		would measure the effectiveness of these controls, e.g.: the application of items in the identifier hierarchy for selection of the specified identifier type.
• Data derivations	(Suggested definition for the HRSS context): The extraction of data items from sources / other items or form of processing in a manner which produces an output of defined integrity without compromising the entity or disclosing the identity of the source.	There are multiple uses of data derivations in national dataset records. Linkage to the source in accordance with the derivation rules assures a degree of authentication and acts as a quality control. For example: in data -flows to SUS, the organisation code 'PCT of residence' is derived from the postcode; patient classification. Patient classifications (ordinary admission, day case or regular day or night admission, etc) areas derived from 'intended management', 'admission method' and 'duration of stay' of the patient. HES makes extensive use of data derivations, e.g.: using the postcode to derive at least 15 of the 25 'geographical' items in the HES Data Dictionary. (See also Section 5.2)
• Indications of data derivation	Indications of derivation processes applied - single or multiple – and where multiple routines existed, which stage in the derivation process had produced the useable output.	The application of multiple derivations might be an indicator of the quality (or otherwise) of the source data used for the derivation). Metadata would be needed to indicate the derivation properties of derived items and their sources, derivation routines and at which point in the data processing the derivation had occurred.
• Possibility for duplicate, confused or merged records	Indications of the presence of identical copies of records in addition to their original; records in which components have become disconnected and attributed randomly across more than one patient, resulting in a loss of the record integrity;, and where data from more than one record have been combined.	Duplicate records can be identified readily by, for example: records that share common identifiers which by definition are unique, such as NHS Numbers. Further checks for duplication can be built in, such as those used by HES, which routinely checks for duplicates using fields such as local patient identifier, postcode and date of birth. National data-flows in some systems such as SUS do routinely contain some duplicate data. The current messaging protocols for sending data to SUS result in duplicate records if a patient's postcode changes during the lifetime of a patient record. An update to replace prime recipient (PCT of residence) by senders using "bulk update" message protocols creates a duplicate record. Therefore it is important that the system includes robust functions for identifying duplicates between and within records, preferably prior to the data loading and certainly within the system.
• Default	Data item or code used to provide a component with a valid structure and format as a substitute where the data item is not known or not available.	This needs to identify where defaults have been used. A distinction can be made between 'legitimate' and other types of defaults, e.g.: defaults considered as 'legitimate' for use, in that they provide a means to populate the field with a valid structure and some indication of their meaning status is, are published with national values in the NHS Data Dictionary. They include code for 'Not Stated' in the national values for 'Ethnic Category' and codes for 'Not Known' or 'Not Given' values for certain codes. However, the national trend is to discourage use of defaults, as these do not substitute information content even if they enable data to flow in automated systems.

Data Quality Metrics	Definition (HRSS)	Considerations for DQ measures
• "Consent Flags"	An indicator that the patient has given their consent for use of their identifiable information for the purposes of the research project. Equally importantly, any indicator that the patient has explicitly declined to consent for (particular aspects of) such use.	The application of 'flags' to indicate patient consent or refusal to allow use of their identifiable information will be vital where the researcher needs to know what data can be used when person- identifiers are needed. The 'flags' will need to indicate types of identification to which the patient has consented and the circumstances in which the consent applies (see "Indication of Identifiers" above).
***Coverage attributes*:**		
• Geographic coverage	A measure of the precision of geographic coordinates used and/or a measure of the geographic coverage of the data set.	This metric is designed to provide confidence levels in the geographic assignment of the data. Geographic assignment in information processing is routinely based upon the underlying demographic data in a patient's record. The patient postcode is associated with a range of business applications in health information such as identifying registered GP practice and responsible commissioner. It, and is also used for supporting complex geographical associations in health analysis and research. The 'National Statistics Postcode Directory' (NSPD) supported by the Office for National Statistics (ONS) and updated every 3 months is the essential tool for postcode look-up and derivation used in automated systems throughout the UK. The HRSSsystem would need to support different versions of the file and link them to the data from applicable time periods. Issues with demographic association are more often related to data problems, such as un-reported changes of address. It will be important therefore that the metadata include context about the quality of the source data.
• Age range	The difference between the required minimum and maximum age limits of the cases.	This metric would assess the reliability of processing used for deriving age within defined age parameters. That information would be informed by the metadata on the processes and data on age or dates used to in derivation. Information would also be required on related parameters used at source in some system processing. F, for example: HES deletes dates of birth earlier than 01/01/1885.
• Sample rate	The measurement of the frequency with which snapshots at a given point in time are taken over a continuous progression (of time).	This will need to specify frequency of sampling as the basis for the measure. Metadata will need to describe the contents of the sample, sample type and information relevant to the sample rate.
• Sample type	Defines the format of a single sampling value (used to test hypothesis by taking a small part of something intended to be representative of the whole) – and measure of usefulness with reference to the research.	Information will be needed on the data content of the sample, and intended purpose and outcome of the research. All types of sampling likely to be used for research should be accessible for HRSS users.
• Gender	Classification of objects approximately corresponding	This needs to identify where data supporting gender is reliably specific or non-specific, -

Data Quality Metrics	Definition (HRSS)	Considerations for DQ measures
	to the two sexes and other gender identifiers.	informed by metadata.
• Condition-specific	State of health applicable to the subject of research.	This item is applicable for a DQ metric in the context of clinical health. Information on healthcare context is important for understanding the outcomes and quality. The metrics should be specified from attributes in the underlying data based on the advice of expert users.
• **Time-related:**	Considerations that relate to how up-to-date the data are with respect to the task for which they are to be used, and time properties inherent within the data content	
• Timeliness	Speed of dissemination of the data - i.e., the lapse of time between the end of a reference period (or a reference date) and dissemination of the data.	The length of time between the point at which the source (live) data were taken and when they are available is a DQ / data integrity performance measure, which should be configurable to the requirements of the researcher. As data processing technology improves, the system could expect to deliver the data at considerably shorter intervals than those used for national data-flows. For reporting to SUS, for instance, most organisations are contractually required currently to submit CDS data around 5 days after the month end. Therefore, so some data in that submission will be at least 4 weeks old. 18-Week Wait reporting requirements in 2008/09 are expected to significantly increase timeliness of submissions to at least weekly reporting (in Scotland the target for providing Scottish Morbidity Record (SMR) data is 6 weeks after discharge/attendance). HES data has traditionally been released around 9 months after the end of the relevant financial year (making the early data in each annual release close to two years old)). The facility to view the data in real time would be a future consideration in the system's evolution.
• Temporal coverage - (starting and ending dates, length of history)	Identification of a span of time and its end-points.	The measure of the success in identifying temporal coverage would need to take account of differing requirements, which might need to reference retrospective data collections. Metadata would be needed for context about the composition, quality and access properties of applicable data collections.
• Frequency of updates	Measurement of the number of occurrences within a given time period when more recent data / information is supplied.	Details should be provided of the set intervals between specified updates as a reference point for variances.
• Last update	The most recent replacement (of data) to reflect any changes since the previous version.	The date of the last update would be provided for the required data set and data. Details of the items subject to change should also be reported.
Coding: (including Clinical Coding)	Coding is the conversion of information on a subject or unit into another form according to a defined set of rules, often for the purpose of data storage. Clinical coding is the assignment of clinical activity in different settings in accordance with a disease and	The DQ metrics for clinical coding would support the requirements for clinical classifications used in the NHS to support reports of clinical activity in national data-flows – specifically diagnoses (disease classifications) and procedures carried out in patient treatment. The national classifications are ICD and OPCS, and Read codes. The system will also need to support Snomed CT when the use of clinical terms is

Data Quality Metrics	Definition (HRSS)	Considerations for DQ measures
	procedure classification system, which permits the systematic recording, analysis, interpretation and comparison of morbidity and procedural data.	enabled in the NHS more widely than at present - currently mainly in pathology.
• Coding scheme	The parameters assigned to the coded values used in the design of specific coding schemes / rules.	Details would be needed on what coding schemes had been used (e.g.: ICD-10, OPCS-4, or Read), and what versions. There are also coding schemes, which are only applicable to certain areas of activity, e.g.: A&E Investigation Codes can only be used in the A&E CDS. Some classification schemes are subject to regular updates, e.g.: NHS Connecting for Health (NHS CFH) Classifications released OPCS-4.4 coding for use in 2007/08 data. Currently in the UK there are tight criteria for national clinical coding schemes that, which are acceptable for statistical purposes e.g.: Read codes (clinical terms) cannot be used for HES without an equivalent ICD-10 or OPCS-4 code.
• Coding validity	The data conform to the defined coding structure.	Validity measures present additional complexities in rules of classifications compared with other areas of patient data. The validation applied in XML messaging for national data-flows to SUS, for instance, will validate on published standard values for many data items, but will not do so currently for clinical classifications because of the complexity of coding values. Therefore it is possible to submit an ICD or OPCS code, which does not exist in published classifications although it has a valid structure. However, some system-based validation can be applied to clinical classifications. For example, system checks can be used to support sequence-based rules, as in ICD-10 where codes in the range V01 – Y98 (external causes) cannot be used in the primary diagnosis field. The validity criteria for clinical coding that would operate at the HRSS would need to be determined to meet the requirements of research.
• Coding completeness	The data have been recorded in all the required coded fields to at least the minimum required specificity.	Similar considerations apply as those for "completeness" metrics in other data – see above.
• Coding depth	Use of the range of available digits in the classification code structure to present degrees of clinical complexity and dimension.	Rules for national classifications support varying levels of coding complexity within a single code. In ICD-10 the rules expect coding to 4th digit, but also support additional 5th and 6th digit complexities (an 'A' or a 'D' in the 6th digit designates an asterisk or dagger indicator and applicable codes will need to be paired or sequenced appropriately). Local requirements for additional detail, if needed for research, may present considerations around national coding standards, e.g.: in OPCS-4 the 5th character is only of local use and is removed in national data-flows.
• Specificity of coding	The application of a clinical code at the most atomic level	It is important for clinicians, and likewise researchers, that the appropriate term or code can

Data Quality Metrics	Definition (HRSS)	Considerations for DQ measures
	for maximum clarity.	be found quickly, accurately, reproducibly, unambiguously, and easily. A coding scheme with greater 'granularity' can better meet the above criteria. It is also more likely that codes with such properties support the business needs of the organisation, e.g. with billing, insurance, reimbursement, clinical decision support, population based care, and the ability to do outcomes research. Above all, it will be a clinician requirement that the diagnosis terminology fully captures the clinical intent. Metadata will need to be present to inform on all the above properties of codes within coding schemes.
• Number of diagnoses codes per entry	Presentation of multiple clinical details.	National data-flows allow for up to 12 repetitions, i.e.: selection of additional codes from the classification scheme. These fields, populated with codes correctly sequenced and assigned, provide an opportunity to record additional clinical dimensions on the case. Payment by Results (PbR) is thought to have acted as an incentive in some NHS organisations to increase the amount of clinical detail recorded and reported for the purpose of driving up the HRG level and the associated tariff.
• Audit of coding	Independent review of coding to assess the application of the classification rules and system controls. Examination and verification of coded data with reference to professional standards, where these exist.	Information would include whether and when the data had been subject to a clinical coding audit in line with national standards for clinical coding audit (maintained by NHS Connecting for Health Classifications). This might indicate whether or not the data in question had been audited, i.e.: included in the audit sample. The metadata would contain audit information such as the audit methodology applied, e.g.: using case notes or an Electronic Patient Record (EPR) where the system is considered to be the definitive source of the record, audit sample size and specification, and the results.
Information about the data quality assessment itself:	Summary details about the DQ assessment carried out supported at the portal.	
When completed / updated	Details would be provided on: dates for completion and update of the assessment; identity of person(s) who completed the assessment; the assessment method used; and review date.	Supporting details of the DQ assessment to support the metric would enable users to rate the value of the assessment for their purposes.
By whom completed / how		
Date for review		

Appendix 2 – Dashboard Functionality

Figure 2 and Figure 3 show examples of a type of product available in the commercial sector - "My Dashboard Portal" - developed by Reliasoft Corporation using Reliability software, which uses multiple mini-charts displayed in a panel configuration for high-level views, and other types of graphic presentations for the detail.

Please note that the contents of the figures that follow are mainly 'captured' from website and the quality of reproduction on screen and in print is therefore degraded. However, the purpose of the Figures is to indicate an approach concerned with layout and content at a high level, as opposed to demonstrating detailed content.

Figure 2. – Dashboard Sample

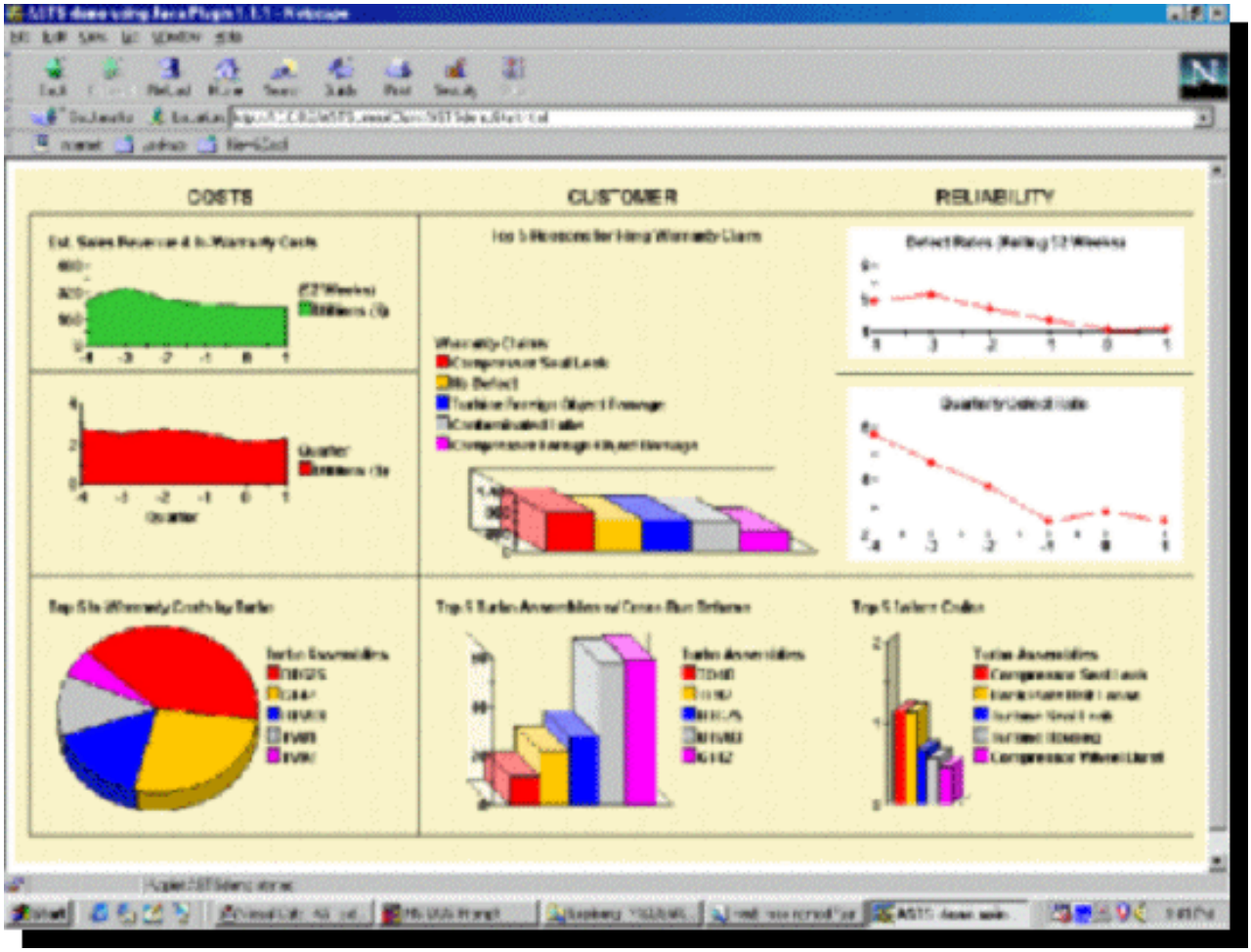
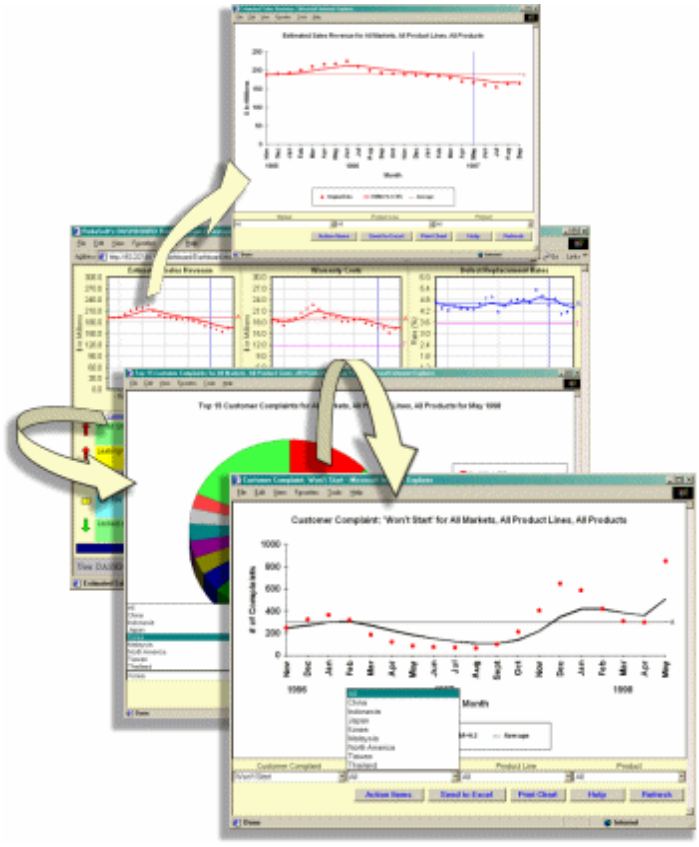


Figure 3. - Dashboard Sample 2



To date there are few known examples of dashboards in the NHS used to display of data quality for reporting on national systems. Examples of dashboards currently in use are described below.

3.3.1. Xcelsius Dashboard

A recent example of a dashboard (– see Figure 4 and 5) & Figure 5 - is an Xcelsius product published in January 2008, which reported on the quality of providers' data on SUS for use in 18-Week Wait reporting. The data quality assessment was carried out on an extract of six months of 2007/08 HES data. It and rated the data items used for 18-Week Wait reporting as compliant or non-compliant with the NHS Data Dictionary standards, compared with the average rating of the selected view. This dashboard can be manipulated to present ratings nationally, by providers in a health authority, or by individual data provider. A commissioning view is currently in development at the Information Centre.

Figure 4. - SUS 18 Weeks Dashboard - 1

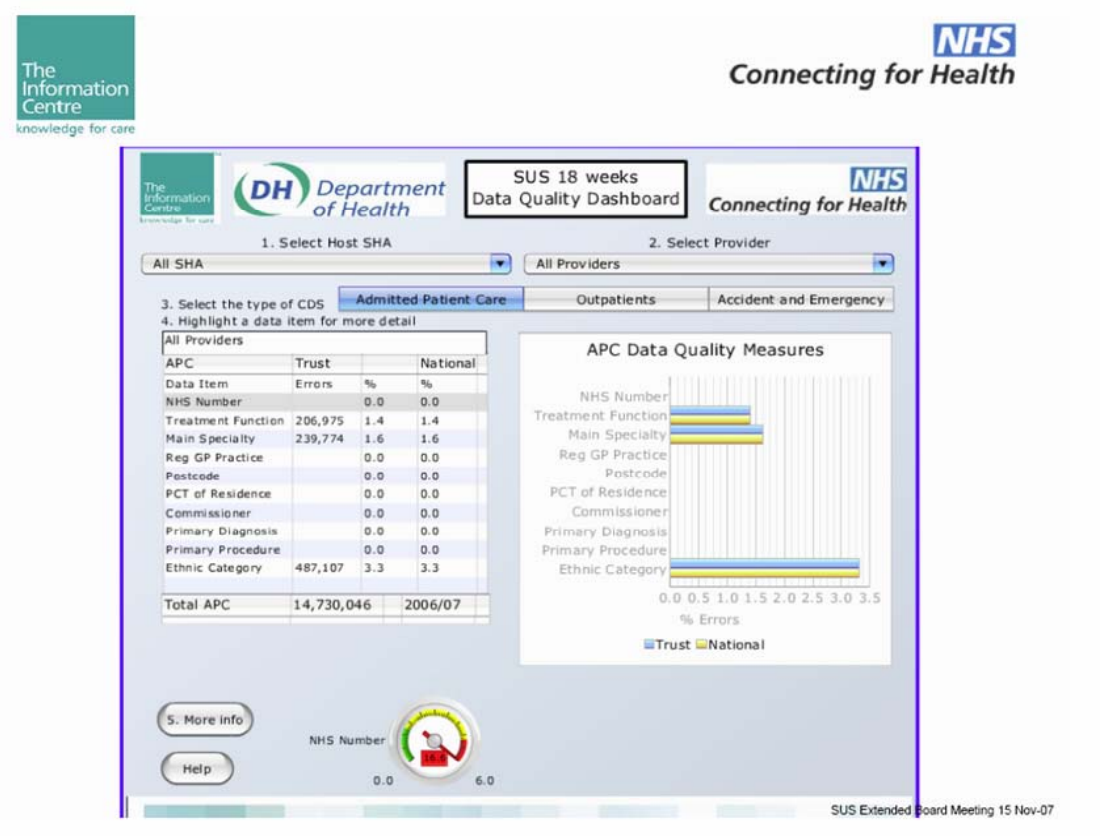


Figure 5. - SUS 18 Weeks Dashboard - 2

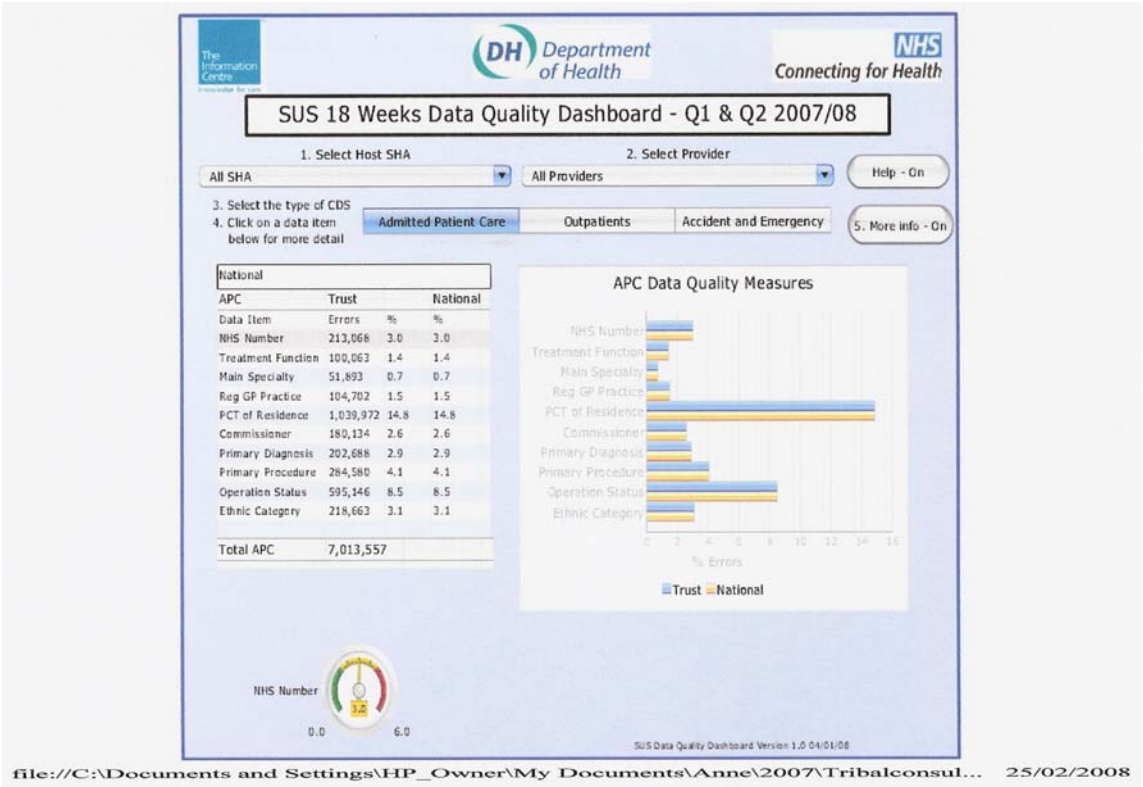
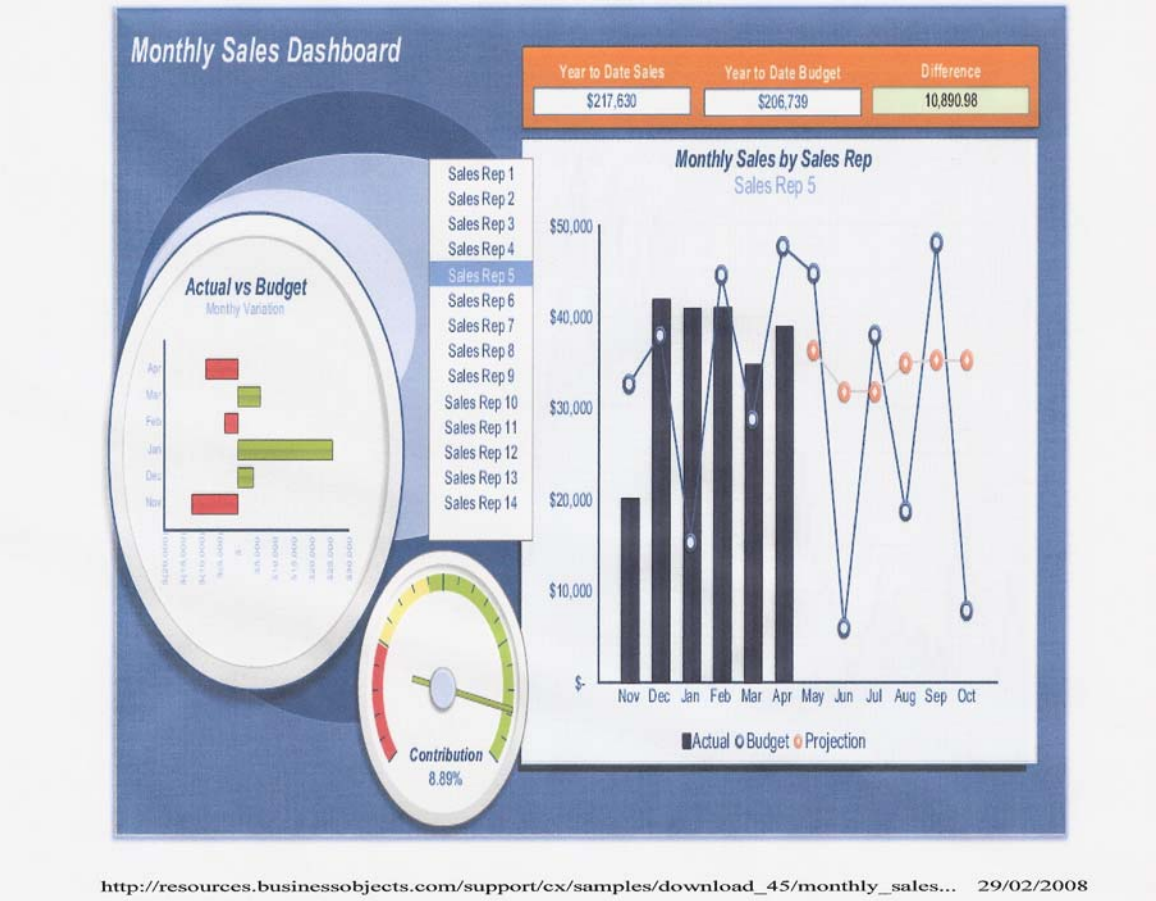


Figure 6. - An example of the Xcelsius product in the commercial sector

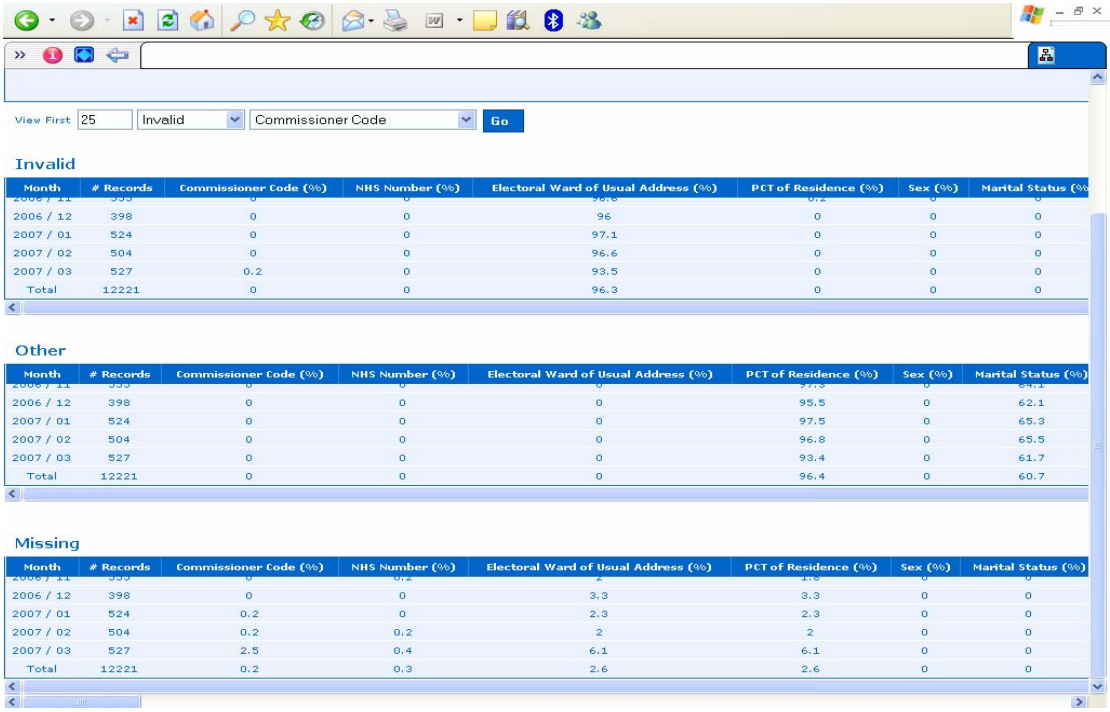


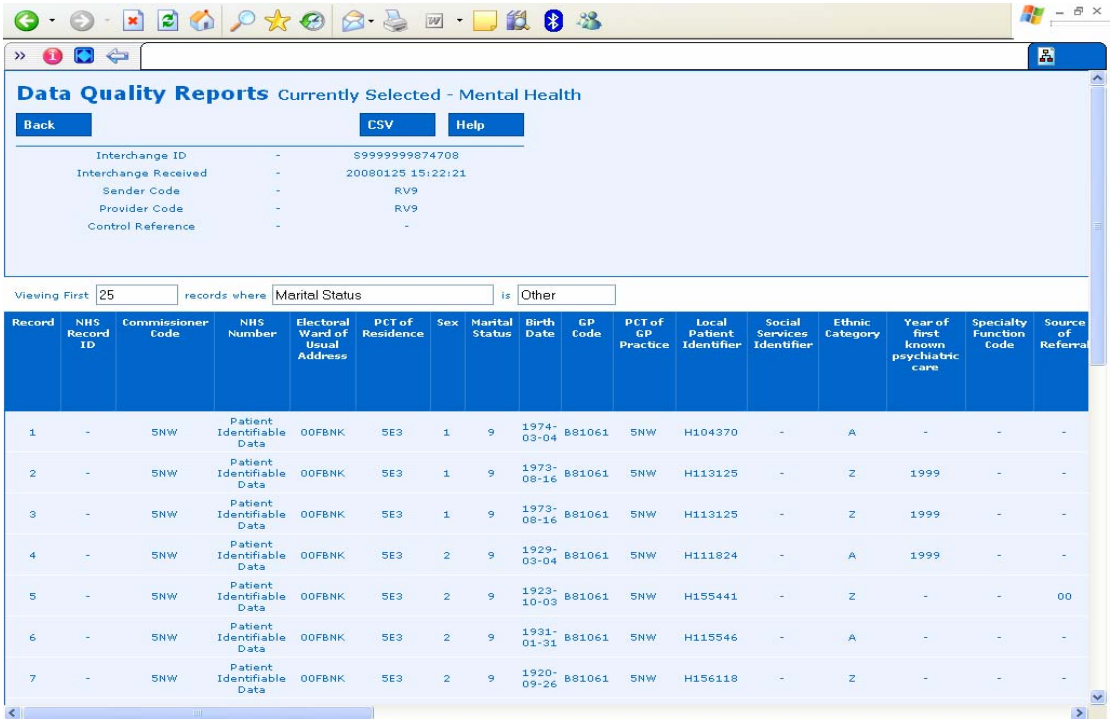
3.3.2. SUS 'Data Quality Report' (DQR)

The 'Data Quality Report' (DQR) on SUS shows the individual data provider the quality of their Commissioning Data Set (CDS) data on SUS and is illustrated in Figure 7. The graphics are limited to Tables showing where errors have occurred at record level. Separate Ttables are provided showing the data quality criteria as 'Valid, Other, Default, ', 'Other', 'Default', 'Invalid' and 'Missing' (VODIM). There is some limited potential for user-configuration, in that providers can drill down further on the summary report to a listing of the first 25 records for each data item which has failed the data quality criteria.

The DQR is also provided jointly with a 'tracker' facility showing providers the progress of loading their CDS submissions into the interchange and, from there, into the SUS database.

Figure 7. - VODIM



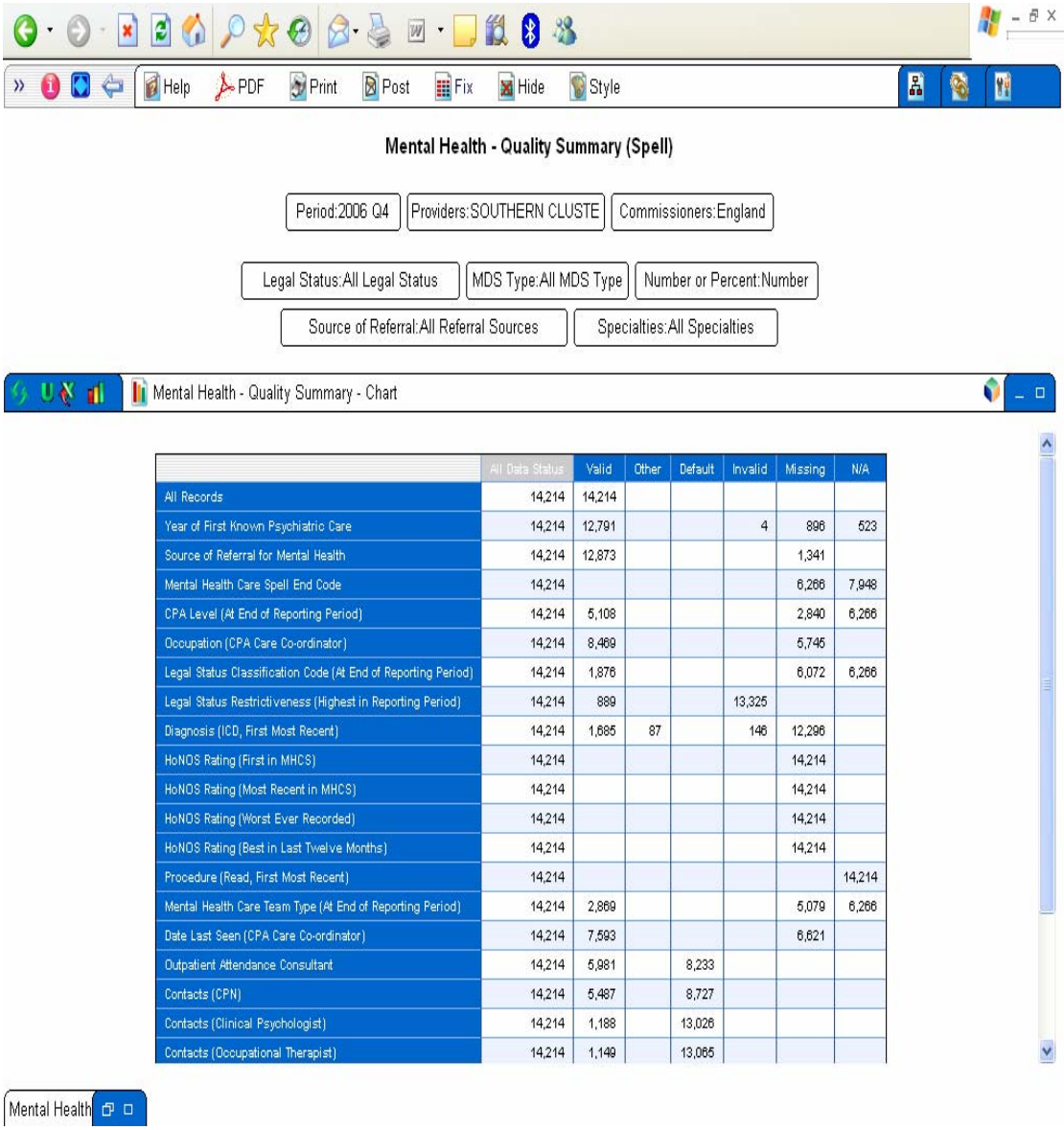


3.3.3. eDQRS on SUS

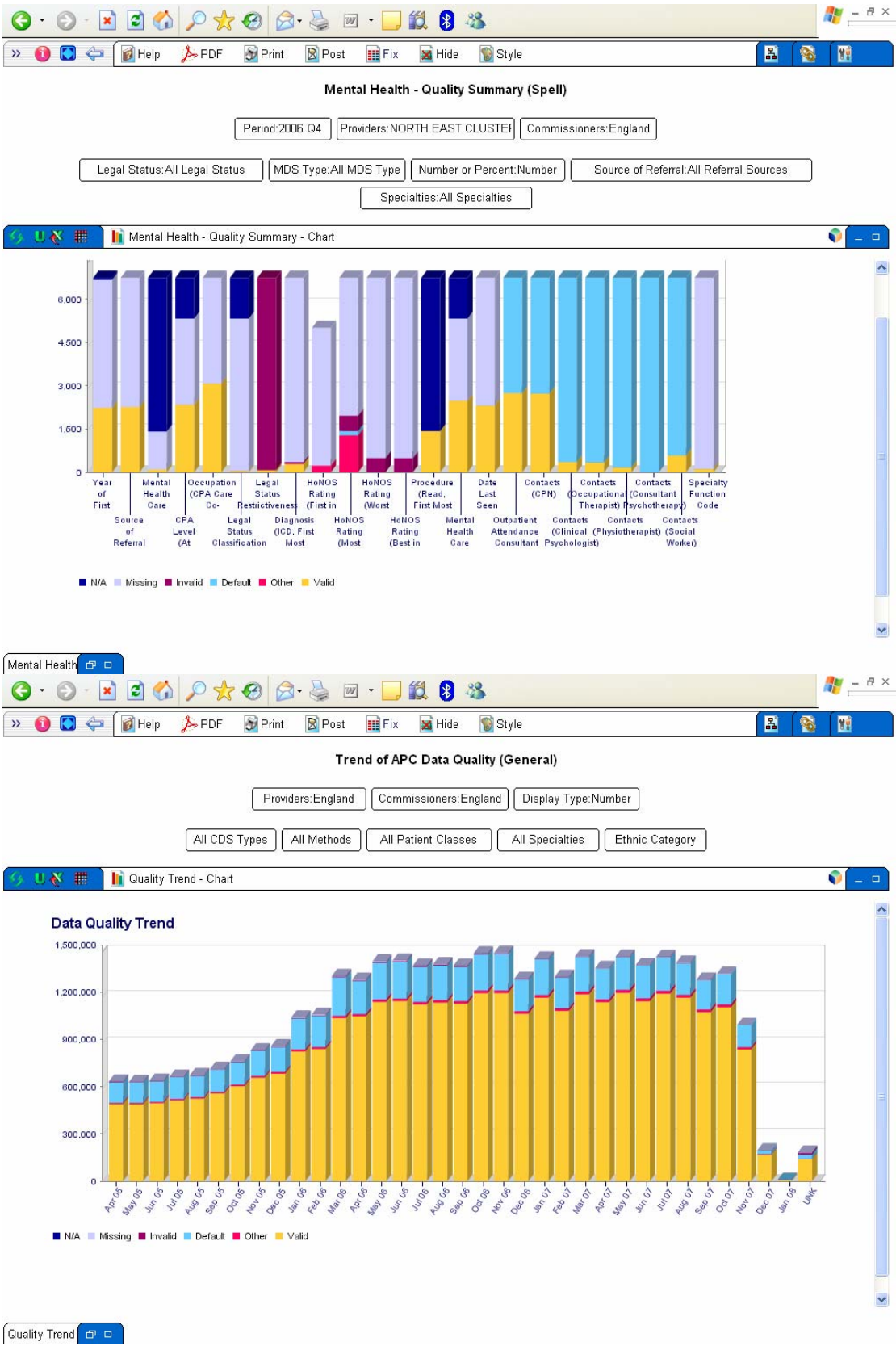
eDQRS is a web-based reporting tool available to all provider and commissioner users of SUS data. Users can configure this tool applying the same 'VODIM' metrics as for the DQR to obtain a view by national, health authority or individual commissioner or provider on the data quality performance of selected data items in the SUS CDS of their choice. Currently eDQRS can report on 34 selected data items in the admitted patient, around 16 in the outpatient and the A&E commissioning datasets, and 35 items in the mental health minimum dataset.

A screenshot of an eDQRS report of data from mental health providers in an English region is shown in Figure 8. A graphical presentation of those data and an example of a national data quality report using eDQRS are shown on the next page.

Figure 8. – eDQRS



	All Data Status	Valid	Other	Default	Invalid	Missing	N/A
All Records	14,214	14,214					
Year of First Known Psychiatric Care	14,214	12,791			4	896	523
Source of Referral for Mental Health	14,214	12,873				1,341	
Mental Health Care Spell End Code	14,214					6,266	7,948
CPA Level (At End of Reporting Period)	14,214	5,108				2,840	6,266
Occupation (CPA Care Co-ordinator)	14,214	8,469				5,745	
Legal Status Classification Code (At End of Reporting Period)	14,214	1,876				6,072	6,266
Legal Status Restrictiveness (Highest in Reporting Period)	14,214	889			13,325		
Diagnosis (ICD, First Most Recent)	14,214	1,685	87		146	12,296	
HoNOS Rating (First in MHCS)	14,214					14,214	
HoNOS Rating (Most Recent in MHCS)	14,214					14,214	
HoNOS Rating (Worst Ever Recorded)	14,214					14,214	
HoNOS Rating (Best in Last Twelve Months)	14,214					14,214	
Procedure (Read, First Most Recent)	14,214						14,214
Mental Health Care Team Type (At End of Reporting Period)	14,214	2,869				5,079	6,266
Date Last Seen (CPA Care Co-ordinator)	14,214	7,593				6,621	
Outpatient Attendance Consultant	14,214	5,981		8,233			
Contacts (CPN)	14,214	5,487		8,727			
Contacts (Clinical Psychologist)	14,214	1,188		13,026			
Contacts (Occupational Therapist)	14,214	1,149		13,065			


Help
PDF
Print
Post
Fix
Hide
Style
Mental Health - Quality Summary (Spell)
Period:2006 Q4
Providers:NORTH EAST CLUSTEI
Commissioners:England
Legal Status:All Legal Status
MDS Type:All MDS Type
Number or Percent:Number
Source of Referral:All Referral Sources
Specialties:All Specialties
Mental Health - Quality Summary - Chart
6,000
4,500
3,000
1,500
0
Year of First
Mental Health Care
Occupation (CPA Care Co-
Legal Status Restrictiveness
HoNOS Rating (First in
HoNOS Rating (Worst
Procedure (Read, First Most
Date Last Seen
Contacts (CPN)
Contacts (Occupational Therapist)
Contacts (Consultant Psychotherapy)
Specialty Function Code
Source of Referral
CPA Level (At
Legal Status Classification
Diagnosis (ICD, First Most
HoNOS Rating (Most
HoNOS Rating (Best in
Mental Health Care
Outpatient Attendance Consultant
Contacts (Clinical Psychologist)
Contacts (Physiotherapist)
Contacts (Social Worker)
N/A
Missing
Invalid
Default
Other
Valid
Mental Health
Help
PDF
Print
Post
Fix
Hide
Style
Trend of APC Data Quality (General)
Providers:England
Commissioners:England
Display Type:Number
All CDS Types
All Methods
All Patient Classes
All Specialties
Ethnic Category
Quality Trend - Chart
Data Quality Trend
1,500,000
1,200,000
900,000
600,000
300,000
0
Apr 05
May 05
Jun 05
Jul 05
Aug 05
Sep 05
Oct 05
Nov 05
Dec 05
Jan 06
Feb 06
Mar 06
Apr 06
May 06
Jun 06
Jul 06
Aug 06
Sep 06
Oct 06
Nov 06
Dec 06
Jan 07
Feb 07
Mar 07
Apr 07
May 07
Jun 07
Jul 07
Aug 07
Sep 07
Oct 07
Nov 07
Dec 07
Jan 08
UNK
N/A
Missing
Invalid
Default
Other
Valid
Quality Trend