Cookbook for Regional Interoperability Detailed Design Paper #026

Data Normalisation

PRELIMINARY DRAFT

Version 1.0 – 9th June 2019

Abstract Interoperability Cookbook Anchor Points

Section	Title
6	FHIR Resource Profiles

Table of Contents

Version Control

Version	Release Date	Released By	Reason for Release
1.0	9/06/2019	R Hickingbotham	Preliminary draft

Reviewers

Initials	Name	Role	Organisation

1 Introduction

1.1 Purpose of this Document

This document is one of a series of design papers which underpin the Abstract of a Cookbook for Regional Interoperability (the Abstract Cookbook). These papers, in their totality, describe the technical components and the standards which form the YHCR System of Systems. They are intended as a basis for developing or procuring software and so are expressed at a level of precision which aims to avoid ambiguity but consequentially, they are focussed to technical readers.

Design papers are anchored to topics which are discussed in the Abstract Cookbook. They are elaborations of the concepts which were first introduced by the abstract and new content is further detail rather than variations of previously established core principles.

This document (design paper 026 - "Data Normalisation") is a statement of Yorkshire & Humber's approach to achieving the data content for a high-quality longitudinal care record. NHS England has been clear that it considers a successful outcome for a LHCR to be a normalised record which spans the population of the LHCR and complies with the Professional Records Standards Body's (PRSB) headings for a longitudinal care record. The record should exist in both personally identifiable form for direct care and in a pseudonymised form for secondary use. The YHCR supports NHS England's objective and this paper sets out the YHCR approach to achieving it

1.2 What is Normalisation?

Normalisation is a direct synonym of standardisation but is used in the context of the LHCRs to imply standardisation of all aspects of data. The scope of normalisation includes:

- the format in which data is presented;
- the use of coding systems to record concepts rather than free text;
- the use of single coding system across all contributors to record a given class of concepts such as diagnoses or drugs;
- consistent interpretation of codes across contributors in representing the same concept;
- completeness of data;
- consistent definition of what data is captured at various stages of a clinical process;
- no duplication of data;
- a clear demarcation between data which pertains to active and historic conditions;
- the presence of causal relationships between data items.

The definition of the longitudinal care record is an aggregation of data from all sources within the YHCR which is normalised according to these principles. The normalisation process removes the inconsistencies which inevitably arise from data being captured in different systems, by different people for different proposes.

Some people also include some more complex ideas under the banner of normalisation. These include:

- grouping codes for analytical purposes (for example grouping all SNOMED-CT codes which imply a diagnosis of a liver disease);
- deriving new data from a cluster of existing data items (for example generating a NEWS 2 score from vital sign observations).

These ideas are included the scope of this paper.

1.3 Complexities in Normalising Data

Normalisation is challenging. Some of the difficulties include:

- Much of the medical record is in free text form. Well normalised data should represent concepts as structured codes. Semantic analysis and machine learning software can offer suggested coding for clinical concepts embedded in the text, but the reliability of automated coding is questionable particularly if it will be the basis of a clinical decision.
- Different care settings use different coding system for the same clinical concepts. Often coding systems do not map onto each other conformally.
- Even when the same coding system is used then different clinicians may choose to use codes differently.
- Some categories of data such as discharge summaries, care plans, and test results are widely distributed across the region and multiple copies of the data exists.
- What constitutes duplicated data to some are important distinct facts to others. Consider for example an allergy intolerance recorded separately by two clinicians. The single fact about the patient may be that he has an allergy intolerance but for some uses knowing that there is corroboratory evidence is important.
- Workflow varies between care settings and the data which is captured at different stages of a workflow is different.
- Cause and effect is often not obvious in a data items. For instance, that a patient received physiotherapy because of pervious attendance for an injury at ED may not be explicit in the data.

The difficulties in normalising data must not detract from the YHCR's ambition. But how the journey is undertaken must reflect practicalities and focus on prioritising normalisation where there is greatest benefit to do so. Clinicians providing direct care are often able to resolve inconsistencies in data: free text is meaningful; they know the practises of other care settings and individuals; and seeing data in the raw form that is was captured is often desirable. The greatest beneficiary of well normalised data is the algorithm: software expects precision in data. But, depending on the purpose for which it is being used, software can be more forgiving of mistakes in data than clinicians using it for clinical decisions. For instance, an algorithm working in a probabilistic manner over large populations to identify correlations in conditions, lifestyles, treatments etc. does not need 100% accuracy to identify a trend or suggest the possibility of an intervention. Where there is tolerance to imperfection then automated normalisation techniques can be more safely employed.

For the purpose of direct care, the early day focus of the YHCR is to improve access to data. This will require standardising the format in which data is moved around the region, but less emphasis is needed on achieving semantic equivalence, at least in the short term.

For secondary use purposes standardising the coding of data is much more important. But there is more tolerance for error: software can be used interpret data, and slight mismatches in coding systems can be overlooked. There would be a loss of resolution in the data, but this is not as important for population health management as it might be for direct care.

1.4 Relationship of this Document with Other Standards

This paper is a statement of intent rather than a design and does not rely on any particular standard although many standards will be used in the implementation of the intent:

- FHIR;
- SNOMED-CT;
- DM+D;
- ICD10;
- LOINC;
- Read Codes.

1.5 Intended Users of the This Document

Anyone with an interest in the YHCR.

2 Normalisation in the YHCR

The YHCR comprises three classes of actors:

- data providers;
- regional infrastructure;
- data consumers.

All have a role in the normalisation of data. However, the intention of the YHCR is to achieve well normalised data as close to its source as possible. Particular emphasis is placed on data providers improving the quality of their data whilst recognising that this will take time. Initially, data consumers will need to accommodate differences in data that ultimately will converge to a standard. Recognition of this fact dictates how consumers will interact with data providers. The region will assist in mediating these relationships by aligning the normalisation capability of a data provider to published maturity model.

2.1 A Shared Responsibility

2.1.1 Data Providers

Data providers are responsible for presenting high quality data at their boundary in a regionally standardised format (FHIR). Ideally, data will only be presented by an organisation or system if the data content is original to that organisation or system. Initially, data providers may choose to present content as it resides natively in local systems, using local coding vocabularies and free text.

Over time data providers, should present data applying regionally approved coding systems in accordance with regionally issued guidance on their usage. This may involve replacing or upgrading local systems or, a data provider may choose to translate data at its boundary from local coding into a regional coding. There are likely to be tactical and strategic options at play.

Ultimately the boundary of an organisation will serve data on-demand which:

- is packaged as FHIR resources and are compatible with the YHCR FHIR Profile (a derivation from the Care Connect profiles and with profiles for non-Care Connect FHIR resource types);
- is semantically equivalent to data items served by other data providers in that it uses regional coding systems in accordance with regional guidance;
- fully covers the medical record as is known to the organisation or system but excludes data items whose provenance is known to derive to another participant in the YHCR;
- uses a regionally modelled workflow (in the presentation of data if not as a model for local business processes) to locate data items to stages in a clinical process;
- identifies data as being current or historic.

In transitioning to the ultimate goal, the data provider will provide meta data which describes the quality of data included in search requests. Meta data will identify:

- gaps in data;
- known quality issues.

2.1.2 Regional Infrastructure

Regional infrastructure is responsible for unifying certain physical concepts on which all data providers should agree. Regional infrastructure will translate references to these concepts in locally served resources to a regionally held golden record. Initially a regional golden record will exist for:

- Patients;
- Practitioners;
- Organisations;
- Locations.

Medications (the drug definition rather than a prescription, dispensation, or administration event) may also be included at a future stage.

Apart from re-referencing data to regionally held resources, central infrastructure will not modify data in transit from a data provider to a data consumer.

Regional Infrastructure will host algorithms which derive data. These subscribe to data which originates locally, interpret it, and create new data. Examples may include:

- scoring of condition acuity (NEWS2, frailty);
- pre-emptive diagnosis (Cancer risk scores, diabetes indicators).

Regionally hosted services, such as terminology translation tools, will assist data providers and consumers in to perform local normalisation activities.

2.1.3 Data Consumers

Data consumers are responsible for interpreting data in the context in which they operate. This may mean:

- removing redundant data;
- classifying and presenting data in appropriate manner;
- informing users of missing data or operating safely in known issues with the quality of data.

Each data consumer is designed for a purpose and it is best qualified to define its own responsibilities in these regards, However, where the consumer is acting as a presentation layer for the purpose of direct care then it is anticipated that the consumer will, initially, present all data that it receives and clearly identify its provenance. As the level of normalisation improves at source then user interfaces may move to aggregating certain data items and presenting data as though it had been sourced from a single longitudinal record.

2.2 De-Duplication

Of all the normalisation topics, de-duplication tends to receive most attention. Much data is duplicated across care settings and NHS England's stated objective for the LHCRE programme is to achieve a longitudinal care record which removes duplicates. The YHCR approach is as follows:

De-duplicate at source where possible – data should only be served from the data provider which created the data. Pathology test results will only be served from the pathology system, not the GP record which received them; Cara plans will be served by the organisation which authored them not other contributing organisations. Discharge summaries will be served from the trust which created them, not by the organisation to which they were sent.

Consolidate common concepts whilst data is in flight – regional infrastructure manages golden record for patients, organisations, practitioners, and location. Local resources and references are substituted for regional ones while data is in transit.

Further de-deduplicate where data lands as is appropriate given context – local data consumers can best determine what constitutes duplicate data given the function that they perform. A data consumer which is attempting to correlate problems being treated with drugs prescribed may which to treat similar diagnoses as the same whereas the distinction may be important for a clinician reviewing the development of a condition.

2.3 Current Problems, Allergies and Medications

Distinguishing current state from data about historic care provision is important for both direct care and secondary uses. However, for a number of reasons different care setting often disagree about what data is current.

There would be a clear benefit to its users for the YHCR to be able to, unambiguously, present certain facts about a patient as being current and this topic has been the subject of considerable debate among the LHCREs, NHS England, and NHS Digital. One position holds that to do this a dataset must be maintained centrally by the LHCR which describes current problems, allergies, and medications. A counter-position argues that such a dataset would only another source of conflicting data.

In consensus with the other LHCREs the YHCR will not attempt to reconcile the different opinions of different care settings as to what constitutes current facts. Nor will it maintain a central dataset which overrules local care settings' interpretation of current facts. A data consumer executing a query for current problems, allergies, and medications will receive an aggregation of all current facts as known to all data providers. The data consumer is responsible for interpreting this information.

The YHCR aims to improve data quality at source. It will offer a process for clinicians to communicate about observed issues and inconsistencies in data with data providers. It will monitor queries for current facts and automatically inform data providers about differences of opinion.

2.4 Spells of Care and the Causality of Data

It is desirable, but at the moment considered to be impractical, to group patient contact across different care settings within a regionally defined spell of care. There is currently no national or regional definition which would allow local systems to group their activities across care settings as relating to the treatment of a particular condition.

The YHCR will offer regionally standardised definitions of an episode of care and encounters which will allow individual care settings to consistently structure causal relationships in their local data.

As normalisation at source improves and there is a consistent representation of referral events then this position will be reviewed to establish whether a chain of care can be construed.

2.5 Data Architecture Design Authority (DADA)

This regional group will be responsible for defining regional FHIR profiles and reviewing their application across the YHCR. The DADA is seen as being a key mechanism for achieving semantic consistency between data sources.

2.6 Population Health Management as a Data Consumer

Platforms for population health management are data consumers in the YHCR and have same opportunities and responsibilities for normalisation as other data consumers. Standard tooling will be installed on the platform which will enable population health users to perform code translations, analyse and apply automatic coding to free text, and index coding systems in a manner which allows data to be easily searched.

SNOMED-CT presents particular challenges for analytics platforms. The coding system is represented in a graph structure with no natural hierarchy in the concept codes. Grouping detailed coding into a more summary form can be computationally expensive unless innovative indexing techniques are employed.

The sophistication of normalisation at source will evolve over time. Initially there will be an emphasis by data providers on achieving high data coverage and uniformity in the format of data but this will be at the expense of semantic equivalence of data. This priority will necessitate, in the early days, more normalisation work by population health management platforms than would be the ultimate goal.

2.7 Use of the NHS Number in the YHCR

The NHS Number is the only patient identifier used by the YHCR. All patients registered with the YHCR must have an NHS Number. With the exception of new-born babies, a data provider must have traced the NHS number with national Patient Demographic Service (PDS) before releasing data to the YHCR. Data for new-born babies may be released to the YHCR for up to six weeks from the birth date without having a traced NHS Number.

The YHCR will offer search facilities to help data consumers to locate an NHS Number. However, the data consumer is ultimately responsible for ensuring that it has the correct NHS Number and understands the clinical risk of using an NHS Number in which it has less than complete confidence.

2.8 Normalisation and Maturity

The YHCR is taking a pragmatic approach to achieving a well normalised longitudinal care record with good coverage across its services and population.

Early day focus is on enabling data to flow, addressing the most accessible sources of data first and prioritising access to data over semantic standardisation at source. However, the aim is to rapidly improve maturity of which there are a number of different measurements:

- the number and type of organisations contributing data to the YHCR;
- the number of care professionals with access to the YHCR;
- the sophistication of secondary uses of YHCR data;
- bidirectional flows of data enabling cross care setting care co-ordination;
- the types of FHIR resources available from data providers;

- the coverage of data managed by data providers which is mapped to FHIR resources;
- the use of regional coding systems by data providers and the quality of coding.

Measurements which are aligned to the maturity of the organisations contributing to the YHCR are controlled by a maturity model which is the subject of a separate design paper (design paper 023 – "The YHCR Maturity Model". At the time of writing, the model is still being worked upon and the design paper should be treated as the definitive guide. However, salient features which give an indication of its direction are illustrated below:

Maturity Model

		Level 1	Level 2	Level 3	Level 4	Level 5	Level 6
Data Providers	**Breadth of FHIR Resource Coverage**	Demographics. Appointments. Encounters.	Allergy Intolerances. Observations. Care Plans. Conditions.	Case Notes. Clinic Letters. Discharge Summaries.	Test Orders. Medications. Diagnoses. Current Lists.	Questionnaires. Immunizations. Risk Assessments.	Images and Other Media.
	Depth of Data Coverage	Patient or Client Administration Systems.	Primary EPR. Radiology. Pathology. Oncology. ED.	Document Management. Vital Signs. Maternity.	Orders & Results. ePMA.	Less Accessible Departmental Systems.	PACS. Endoscopy.
	Technical Maturity	Direct Resource Access. Basic Searches.	Expanded Search. Subscriptions.	Asynchronous Searching.	Full Searching Acceptance of Inbound Resources.	Patching, Chained Searching.	Local Consent. Peer to Peer Relationships.
	Normalisation At Source	Normalised Data Format (FHIR)	Some Use of SNOMED-CT	Extended Use of SNOMED-CT	Use of DM+D	Full Adherence to Regional Coding Standards	Alignment of Data to Regional Model Business Processses
Data Consumers	**Sophistication of Access for Direct Care**	Regional Encounter Data Accessible by Small Group of Users	Clinical Facts Separated by Data Source	Access to YHCR Rolled-out to Majority of Clinicians	Some Merging of Data from Different Sources	Removal of Redundant Data. Prioritisation of Presentation of Data Items	Presentation of Longitudinal Record as a Single Record
	Sophistication of Secondary Use	YHCR Used a Data Source	Local Codes Translated Into Common Systems	Some Meaning Derived from Free Text	Derived Insight at a Population Level Direct Care Delivery	Care Pathways Analysed for Adherence with Regional Protocols	Derived Insight at an Individual Level Prompts for Intervention in Care
	Bidirectional Data Flows	New Transfer of Care Transactions	Inter Care Setting Referals	Inter Care Setting Care Planning	Inter Care Setting Workflow and Team Working	Automated Execution of Care Pathways	Enablement of Major and Rapid Restructuring of Provision of Care

Appendix 1 – Maturity Matrix

Section	Narrative	Consultative	Draft	Normative
1 Introduction 1.1 Purpose of this Document	X			
1.2 What is Normalisation?	X			
1.3 Complexities in Normalising Data	X			
1.4 Relationship of this Document with Other Standards	X			
1.5 Intended Users of the This Document	X			
2 2 Normalisation in the YHCR 2.1 A Shared Responsibility 2.1.1 Data Providers			X	
2.1.2 Regional Infrastructure			X	
2.1.3 Data Consumers			X	
2.2 De-Duplication			X	
2.3 Current Problems, Allergies and Medications			X	
2.4 Spells of Care and the Causality of Data			X	
2.5 Data Architecture Design Authority (DADA)			X	
2.6 Population Health Management as a Data Consumer			X	
2.7 Use of the NHS Number in the YHCR			X	
2.8 Normalisation and Maturity			X	