Dataset Guidelines

Authors: The Informatics and Audit Committee, Royal College of Ophthalmologists
Version: 1.0
Date issued: 16 August 2013

Amendment Record

Issue	Description	Author	Date
0.8	First Draft	G W Aylward	24 Nov 2011
1.0	First working version	G W Aylward	16 Aug 2013

Table of Contents

Introduction

A dataset comprises a set of defined variables representing clinical information about a patient with a given condition. A formal Royal College interest in ophthalmic datasets began in 2002 with the initiation of work on the cataract national dataset. Under the umbrella of the NSF for diabetes, a diabetic retinopathy screening dataset was developed and subsequently the Do Once And Share (DOAS) program supported further dataset work on cataract, glaucoma and diabetic eye care. The widespread use of the cataract national dataset has facilitated useful national audits.[1]

However, at present datasets only exist for a small number of conditions, and are in active use for only three (Cataract, diabetic retinopathy, and retinal detachment). There is a need to develop datasets for a much wider range of conditions. Such datasets will facilitate clinical care, research, audit, and revalidation. Common use of the dataset will ensure that information collected by different clinicians, using different paper or electronic systems in different locations, is easily transferable, and can therefore form the basis of large, anonymised databases for audit and outcomes research.[2] This document describes guidelines for establishing new datasets.

Structure of Dataset

The purpose of a dataset is to represent an agreed set of clinical information which can be collected on patients with a particular diagnosis. As well as defining the items to be collected, the dataset should also describe the data type for each. Each item should also be categorised according to the scheme in the following table. Each category is by definition a subset of the category below it;

Category	
Revalidation	Required for revalidation purposes
Mandatory	Data items which are essential for all applications, and must be collected
Optional	Data items which are required for some applications, and may be collected

Principles

A dataset should be designed to comply with the following principles;

1. **The dataset should be a subset of information routinely collected**
   The intention is not to burden already busy clinicians with additional work, so the dataset should be constructed of items that are, or should be, recorded as part of the routine clinical management of the patient.

2. **Items not required for likely analysis should be excluded**
   The collection of data requires time and effort, and therefore the total number of items should be kept to a minimum. The range of analyses likely to be conducted on the data is largely predictable, and items not required for these analyses should be excluded.

3. **Items in common with other College datasets should be congruent**
   A number of data items (for example visual acuity, IOP) will be common to other ophthalmic datasets. It makes sense to ensure that only one definition for each item is used throughout all datasets, particularly within a subspecialty.

4. **The dataset should be capable of implementation in an electronic patient record**
   It is likely that the maximum benefit of the dataset will only be achieved when information is being routinely collected using electronic patient record systems. It is therefore essential that it is capable of being implemented electronically.

Datatypes

Each item of the dataset has a data type, which in the majority of cases should be one of the items on the following list. These correspond to data types available in relational database management systems (RDMS), which form the core of most EPR and other systems collecting clinical data.

Type	Description
NULL	A special entity representing an uncertain or unassigned value
INTEGER	An integer value, normally unsigned (i.e. zero or positive values only)
FLOAT	A floating point value, positive or negative

Type	Description
BOOL	A value representing true or false
STRING	A value containing text (alphanumeric data) of unspecified length
ENUM	A value which represents one of a limited range of values
DATE	A value representing a date
DATETIME	A value representing a date and time
LIST	An entity containing one or more values

The term ‘LIST’ is not a data type, but will be used to represent a ‘one-to-many relationship’. This is a standard way in a RDMS of representing data items which can vary in number (for example a patient could have one, two or more symptoms)

Process

The process for construction of a dataset includes the following steps;

1. Identify the need

The need for a new dataset might come from a wide range of sources, including the College, individual Ophthalmologists, researchers, or specialist societies. The College will keep a register of datasets that are in development, to avoid duplication of effort.

2. Establish a working group

The main business of the working group is to agree the fine detail of the items that are to be included, decisions which might be controversial. It is therefore of vital importance that the group is seen to be representative of the potential users, but not so large as to be unmanageable. A group of six to ten members is recommended. Representation should be sought from as wide a range of ophthalmic working environments as possible, reflecting the ‘target audience’ for the dataset, and increasing the sense of ‘ownership’ of the dataset.

Where they exist, appropriate specialist societies should be consulted and collaborated with in the establishment of the working group.

The working group will require a chair who is responsible for drafting the dataset, arranging meetings or the working group, and incorporating changes into the final document. It is likely that the interval between the first meeting and the production of the final draft will be of the order of six months.

3. Facilitation

The working group should aim to have the first meeting face to face, but subsequent meetings could be virtual. A 'list server' email system will allow efficient communication, as well as storage of comments for reference.

4. Testing

Members of the working group should then test the dataset in the working environment, either using paper forms, or preferably an electronic version of the dataset. This phase of the process may result in changes to the dataset.

5. Approval

Once the dataset is completed by the working group, it is submitted to the Informatics and Audit Committee for approval. The approval criteria should include the following;

- Does the working group include appropriate representation?
- Has the appropriate specialist society been consulted and involved?
- Do all the items in the dataset have an approved data type?
- Does the dataset utilise existing data items and types (e.g. visual acuity, intraocular pressure etc)?
- Has it been adequately tested?
- Have the revalidation committee been consulted (so that any items required for revalidation are included)?

6. Publication

The dataset is then published on the College website so that interested parties can download it and consider implementing it within electronic patient records, audit programs and other systems.

7. Enhancements

The use of the dataset in real systems may well reveal the need for enhancements or changes. While not needing to meet regularly, the working group should remain constituted in order to deal with such enhancement requests.

Standard Items

Certain items in datasets will be common to others (e.g. visual acuity). A parallel work stream within the college is to produce a library of such standard items to ease construction of future datasets and facilitate data sharing. An example is patient demographics, which should include the items in the following table;

Item	Description	Values/format
Patient ID	An identifier which will uniquely identify the patient. In England and Wales this could be the NHS number. This would be removed in anonymised datasets, or replaced with a UUID	INTEGER
Age	The age of the patient in years at the time of presentation. Age provides sufficient information for scientific analysis, without also being patient identifiable data (PID), unlike date of birth	INTEGER
Sex	The patient's gender	ENUM (Male, Female)□
Postcode	The postcode district (outward code). This is the first part of a postcode, and generally corresponds to a post town. It gives useful information for demographic analysis, without being PID	STRING
Consultant	Identifier for consultant in charge of patient (to allow individual audits)	INTEGER
Ethnic category	The ethnicity of the patient using the classification used for the 2001 census[3]	ENUM (British, Irish, Any other White background, White and Black Caribbean, White and Black African, White and Asian, Any other mixed background, Indian, Pakistani, Bangladeshi, Any other Asian background, Caribbean, African, Any other Black background, Chinese, Any other ethnic group, Not stated)

Item	Description	Values/format
Route of referral	Route by which patient arrived in the ophthalmic department, based on who made the initial diagnosis (e.g. if an Optometrist sends a patient via the GP with a suspected diagnosis of RRD, this item would have a value of ‘Optometrist’)□	ENUM (Optometrist, GP, Ophthalmologist from other Trust, Ophthalmologist from same Trust, General A&E, Ophthalmic A&E, New diagnosis in clinic, Other)

All data sets should adhere to the standards for the clinical structure and content of patient records were published in July 2013. http://www.rcplondon.ac.uk/resources/standards-clinical-structure-and-content-patient-records

References

1. Cataract National Dataset V1.2 – Royal College of Ophthalmologists.
2. Narendran N, Jaycock P, Johnston RL, Taylor H, Adams M, Tole DM, Asaria RH, Galloway P, Sparrow JM. The Cataract National Dataset electronic multicentre audit of 55 567 operations: risk stratification for posterior capsule rupture and vitreous loss. *Eye* 2009;**23**:31–37.