Essential Cancer Data Guide Importance of essential data items for cancer in 100,000 Genomes Project Guidance

Functional Area	Participants	Document Key	PAR-GUI-047
Document Owner	Amanda O'Neill	**Status**	Final
Document Author	Clare Craig & Amanda O'Neill	**Version**	1.0
Document Reviewer(s)	Tom Fowler, Deputy Chief Scientist	**Version Date**	14 December 17
	Angela Hamblin, Clinical Lead Cancer	**Next Review Date**	
Document Approval	Tom Fowler, Deputy Chief Scientist Sandi Deans, National Laboratory & Scientific Lead		
Electronic Signature	Tom Fowler Tom Fowler (Dec 14, 2017) Sandi Deans Sandi Deans (Dec 14, 2017)	**Approval Date**	Dec 14, 2017 Dec 14, 2017
Impact on Competent Personnel *(please choose Y/N in the boxes to the right)*		**Read and understand** **Re-train**	Y Y
Transaction ID	Adobe Sign Transaction Number: CBJCHBCAABAAbRuUl8ofbxblRQ4yGNLoYHwtKz9PWMHx		

1 Document History and Control

The controlled copy of this document is maintained in the Genomics England internal document management system. Any copies of this document held outside of that system, in whatever format (for example, paper, email attachment), are considered to have passed out of control and should be checked for currency and validity. This document is uncontrolled when printed.

1.1 Version History

Version	Date	Description
0.1	August 2016	First draft by Clare Craig
0.2	17 Oct. 17	Changes from Clare Craig and Amanda O'Neill following review by Tom Fowler and essential data review
0.3	7 Dec 17	Incorporating changes recommended by Sandi Deans in consultation with Clare Craig
1.0	13 Dec 17	Final version for publication. With final amendments from Sandi Deans and Clare Craig

Contents

2 The purpose of this document

The purpose of this document is to help NHS GMC cancer leads, informatics leads and developers to understand the importance of the essential cancer data and the impact of submitted data on the results of sequencing interpretation reports. It describes a number of scenarios and how they would be interpreted. It also describes the hierarchy rules of the interpretation programme, explaining which data items are used in their order of preference.

This document should be read in conjunction with the Genomics England - Cancer Data Model V3: essential data items for returning results and the Cancer Data Model submission guides.

3 Introduction

This document sets out the decision making behind which cancer data fields need to be mandatory. Mandatory data is required for accurate diagnostic reporting.

The systems that are in place may have been set up to provide all data, both mandatory and optional. There are mandatory items within the registration and consent; sample metadata and diagnosis submissions. These can be broadly divided into three groups: identifiers; data for diagnosis and programme data (which facilitates the flow of sequencing, interpretation and results).

4 Identifiers

Identifiers are needed to ensure tracking of samples through the entire pipeline. Different identifiers are used to track at different stages. This means multiple identifiers must be sent with the sample. In addition, there are tiers of identifiers needed to ensure that, where a patient has more than one primary cancer or sample, we are able to link the metadata to the correct sample.

5 Data for diagnosis and reporting

The five key data points required are:

1. Disease type (organ system involved)
2. Disease subtype (which cancer diagnosis)
3. Tumour type (primary, recurrent or metastatic)
4. Topography codes (where in the body the tissue was removed from)
5. Morphology codes (which cancer diagnosis)

5.1 Impact 1: Diagnosis

When analysing the cancer genome, the majority of the diagnostic information can be gleaned for subtracting the germline DNA from the tumour DNA to reveal the somatic mutations that have arisen over the course of the tumour development. In addition, the germline is examined to look for mutations that have contributed to the tumour's growth but have been present in that patient from birth – the pertinent germline findings. It is essential we have the correct diagnosis is provided to ensure the appropriate parts of the genome are examined.

The disease type and subtype data are used to decide which genes should be examined for mutations. These data can be submitted both at registration and within sample metadata.

It is possible for these data to change over the course of the patients' diagnosis and treatment pathway. For example, a patient may be registered as having a primary brain malignancy (adult glioma), however, pathological examination reveals that it is in fact a metastatic breast cancer. Instead of examining the genome for pertinent genes for breast cancer, if the disease type was not updated correctly in the sample metadata then the patient would receive a report on the genes pertinent for adult gliomas.

Validation of the data can help with this. Two levels of validation are required. Once the data has been received validation can be carried out between data fields that relate to one another.

For example, where the tumour type is primary it is possible to validate that the disease type and topography codes refer to the same part of the body. In contrast, where the tumour type is metastatic it is possible to validate that the disease type and topography codes are from different parts of the body. It is also possible to validate that the disease subtype and morphology codes concur. Where the disease subtype is entered as "other" it is absolutely critical that an accurate morphology coding is submitted to provide sufficient information for diagnosis.

The second level of validation needed is clinical validation at NHS GMC level. There are circumstances where a malignancy can metastasise within the same organ system e.g. malignant melanoma to elsewhere in the skin or lung cancer to elsewhere in the lung. In these instances, there is no mechanism for identifying that the disease type was correct unless a clinician could provide validation.

In addition, if the tumour type was incorrectly entered an inappropriate tumour panel would be applied. For example, a patient was thought to have lung cancer but on pathological diagnosis it was revealed to be a metastatic adenocarcinoma of the colon. In this scenario, the topography code would state lung and the morphology code may state adenocarcinoma. If the tumour type was entered incorrectly as primary, then lung cancer mutations would be searched for rather than colorectal mutations.

5.2 Impact 2: reporting

The variants reported in the analysis are ranked according to their relevance to that disease type therefore it is important that the disease type is correct. Where it is incorrect then the variant presented in tier one will not be those most clinically applicable for that patient's cancer. The numerous tier two variants would need to be searched through in order to identify those relevant to the patient.

6 Programme data

The remaining data points are requested for the following reasons:

1. To ensure there is a record of valid consent from the patient.
2. Gender is used to validate that the correct germline sequence has been applied.
3. The *'responsible consultant'* data item is used to ensure the report is returned to the clinician treating the patient.
4. The diagnosis xml data is used as a final level of validation before reporting.
5. The *'previous treatment'* data item is needed to produce information on the disruption of the genome caused by treatment.
6. The *'tumour content'* is needed to indicate the degree of sensitivity in the interpretation. In certain haematological cancers, it may not be possible to provide

us with a pathological assessment of tumour content; in these cases, NHS GMCs should submit a clinical assessment of tumour content.

7. When a stored sample is sent it should be noted in the 'retrospective sample' data point to ensure these cases do not adversely affect the turnaround time metrics, and we can ensure the quality of these samples is up to standard.
8. Sample handling data points are requested to enable learning about the samples submitted and develop advice on optimal sampling. These data points include: *'tissue source'; 'number of biopsies'; 'macrodissected', 'snap freezing start datetime'; 'tumour type'; 'scroll thickness'; 'number of scrolls'; 'number of sections'; 'section thickness'*; and formalin process fields for formalin fixed samples. It is planned to relax the validation on these fields to in effect make their submission optional. This will not change the message format or the underlying data model. We will inform NHS GMCs when these changes are made, so they can choose to relax the validation in their local systems. The fields that will have their validation relaxed are indicated in the Genomics England - Cancer Data Model V3: essential data items for returning results.

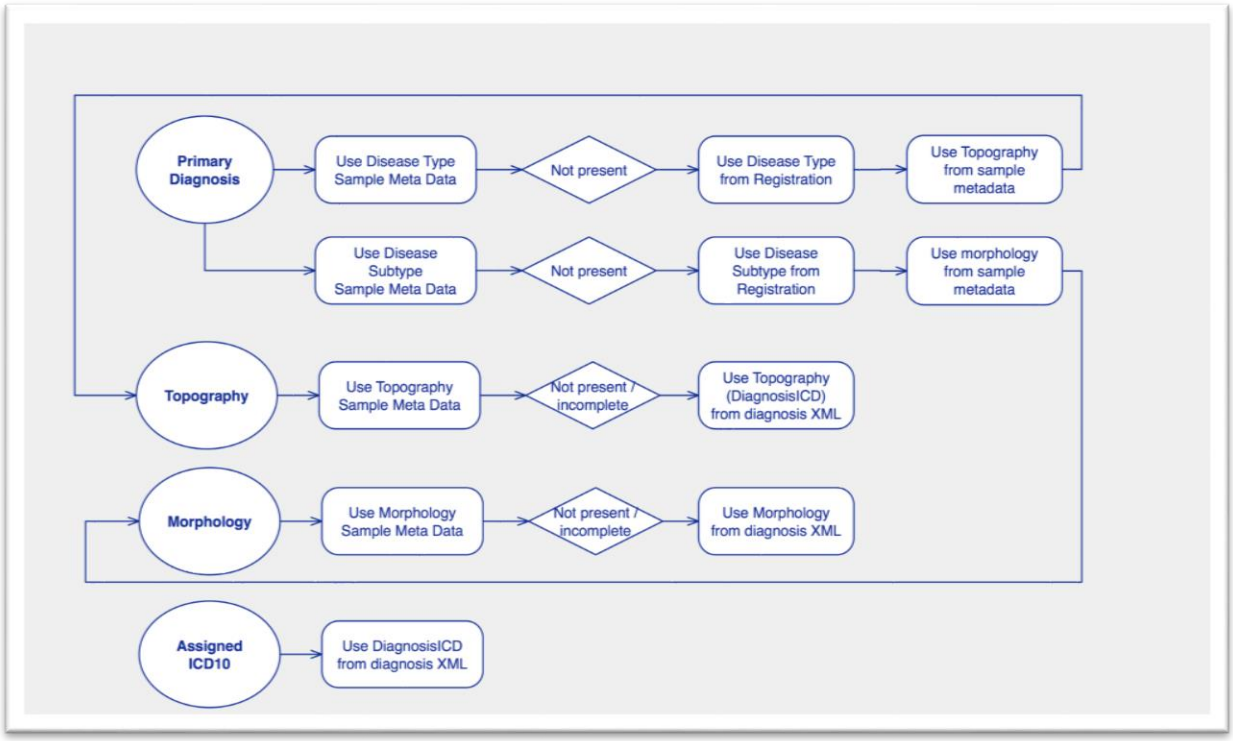
7 Fast track

The priority with fast track data is speed to ensure a timely report. In some circumstances, this might mean that the data will need to be corrected in due course once a more accurate clinical diagnosis has been reached. In such a scenario, where the clinical data is updated over time, then an updated report may need to be issued. The diagnosis xml is critical to allowing us to validate the sample metadata, even if the first report has been issued prior to the submission of the diagnosis xml.

8 Hierarchy of rules

In the interpretation of results process we use the data supplied in the essential cancer data in combination to determine the correct primary cancer diagnosis. This process is based on a hierarchy of the use of data fields to identify the correct clinical data on which we should base the interpretation. This is described as a flow chart on the key diagnosis data points.

8.1 Figure 1: Clinical data hierarchy logic flow for returning results



8.2 Query Resolution

Genomics England have established a query resolution programme to resolve issues in the essential data for the return of results. A series of rules are run on the submitted data, using the hierarchy above and queries are produced where there are gaps or inconsistencies on the data. NHS GMCs will be sent the results of these validation rules on a regular basis and NHS GMCs will be asked to fix these gaps by resubmitting data. The validation rules and methodology for fixing the queries will change over time. Therefore, the current rules and methodology to fix them will be included in the query resolution reports and published on the Genomics England service desk self-help pages.