


SOP: Data Validation Plan for 1k Gene Consortium

SOP Number: DQ-SOP-004/ Version 1.0

| | | | |
|--------------------------------------|---|------------------------------|-----------|
| SOP Number | DQ-SOP-004 | Review Date | 01-DEC-16 |
| Document Owner | Ade Okelarín, Senior Clinical Data Manager | Version | 1.0 |
| Document Author | Rob Jackson, Clinical Data Manager | Version Date | 15-NOV-16 |
| Document Reviewers | Clare Craig, Lead Cancer Clinician Alona Sosinsky, Lead Cancer Bioinformatician | | |
| Document Approval | Amanda O'Neill, Director Clinical Data | Status | FINAL |
| Signature |  | Approval Date | 15 Nov 16 |
| Impact on Competent Personnel | | Read & Understand | X |
| | | Re-train | |

The Master Document containing original signatures is filed by the Data Quality Department. PDF Copies of these procedures are available electronically in the document repository:

<https://my.huddle.net/workspace/38031841/files/#/folder/42474551/list>

Printed copies constitute uncontrolled documents and it is the responsibility of the user to ensure for quality & compliance reasons that any printed SOP reflects the current version held in the document repository

Data validation will be conducted to ensure that data quality can be assured before external releases of data.

3.2. Process Flow & Procedural Steps

There are two distinct data validation processes:

1) Reconciliation of external datasets

Cancer Outcomes and Services Dataset (COSD) will be obtained from Public Health England (PHE) along with Hospital Episode Statistics (HES) from NHS Digital. Extracts will be moved to a secure location (to be defined)

Datapoints that are collected in the external datasets and that are also collected in the GELDAM dataset will be identified, reported and compared by the CDM.

Any discrepancies found will result in DC requests to the appropriate GMC as per DQ-SOP-005. CDM/CDO will ensure these amendments have been correctly applied after future subsequent downloads of external datasets into GEL. Amendments to be made to the GELDAM dataset will be handled as per DQ-SOP-005.

2) Data Reports

The CDM will generate the following reports:

All data records in the GELDAM dataset that fail the validations as per table below.

All discrepant records within the GELDAM dataset as per table below.

Any missing data identified within the table below.

Data reports will be stored in a restricted access area on Huddle.

| DATA MODEL NAME | CODE | LK FIELD NAME | LK VIEW | MANDATORY | VALIDATION RULES | ADDITIONAL REVIEW |
|-----------------------|---------|---------------------------|------------------------------------|---|--|---|
| Topography (SnomedRT) | 31227.1 | Topography (SNOMED RT) | Sample tracking - Collected sample | | Code validity check against online database. ≠ M09350 (morphological description only); ≠ M09450 (no evidence of malignancy) | |
| Topography (SnomedCT) | 14876.2 | Topography (SNOMED CT) | Sample tracking - Collected sample | | Code validity check against online database. ≠ M09350 (morphological description only); ≠ M09450 (no evidence of malignancy) | Produce listing where Topography (SNOMED CT) ≠ Diagnosis (SNOMED CT), per GMC |
| Metastatic Site | 14937.1 | Metastatic Site Id | Diagnosis | No | <i>valid enumeration as per data model</i> | Frequency of '07' (unknown) and '99' (other) per GMC |
| Integrated T | 14944.1 | Composite Tnm Integratedt | Diagnosis | If Disease type ≠ 'Haematological'; 'Ovarian'; 'Endometrial'; 'Colorectal'; 'Malignant melanoma' | Format (where <i>n</i> = any number; <i>a</i> = any letter): = "T" <i>n</i> = "T" <i>n</i> <i>a</i> = "TA" ≠ "Tx" ≠ "T" | |
| Integrated N | 14945.1 | Composite Tnm Integratedn | Diagnosis | If Disease type ≠ 'Haematological'; 'Ovarian'; 'Endometrial'; 'Colorectal'; 'Malignant melanoma' | Format (where <i>n</i> = any number; <i>a</i> = any letter): = "N" <i>n</i> = "N" <i>n</i> <i>a</i> = "NA" ≠ "Nx" ≠ "N" | |
| Integrated M | 14946.1 | Composite Tnm Integratedm | Diagnosis | If Disease type ≠ 'Haematological'; 'Ovarian'; 'Endometrial'; 'Colorectal'; 'Malignant melanoma' | Format (where <i>n</i> = any number; <i>a</i> = any letter): = "M" <i>n</i> = "M" <i>n</i> <i>a</i> = "MA" ≠ "Mx" ≠ "M" | |

| DATA MODEL NAME | CODE | LK FIELD NAME | LK VIEW | MANDATORY | VALIDATION RULES | ADDITIONAL REVIEW |
|---------------------------------|---------|---------------------------------|-----------|---|---|-------------------|
| Final Figo Stage Version | 33088.1 | Final Figo Stage Version | Diagnosis | If Final Figo Stage ≠ <i>blank</i> | No validation | |
| Furhman Grade | 33070.1 | Furhman Grade | Diagnosis | If Grade of differentiation = GX & If Disease Type = 'Renal' | <i>valid enumeration as per data model</i> | |
| Gleason Grade (Primary) | 33071.1 | Gleason Grade Primary | Diagnosis | If Grade of differentiation = GX & If Disease Type = 'Prostate' | <i>valid enumeration as per data model</i> | |
| Gleason Grade (Secondary) | 33512.1 | Gleason Grade Secondary | Diagnosis | If Gleason Grade (Primary) ≠ <i>blank</i> | <i>valid enumeration as per data model</i> | |
| Figo Grade | 33065.1 | Figo Grade | Diagnosis | If Grade of differentiation = GX & If Disease Type = 'Ovarian'; 'Endometrial' | <i>valid enumeration as per data model</i> | |
| MMR Lynch Mutation Tumour Grade | 33059.1 | MMR Lynch Mutation Tumour Grade | Diagnosis | No | No validation as not being exported to GENE | |
| Tumour Grade (Lung) | 33060.1 | Tumour Grade Lung | Diagnosis | If Grade of differentiation = GX & If Disease Type = 'Lung' | <i>valid enumeration as per data model</i> | |
| Invasive Grade (Breast) | 33062.1 | Invasive Grade Breast | Diagnosis | If Grade of differentiation = GX | <i>valid enumeration as per data model</i> | |

| DATA MODEL NAME | CODE | LK FIELD NAME | LK VIEW | MANDATORY | VALIDATION RULES | ADDITIONAL REVIEW |
|------------------------------|---------|-------------------|------------------------------------|---|---|--|
| Tumour Type | 14721.1 | Tumour type | Sample tracking - Collected sample | Yes | <i>valid enumeration as per data model</i> | |
| Excision Margin | 14904.1 | Excision margin | Sample tracking - Collected sample | Yes | <i>valid enumeration as per data model</i> | Frequency of 98 (Not applicable) and 99 (Not known) |
| Tumour Size | 29075.2 | Tumour size | Sample tracking - Collected sample | Yes | > 0.2 < 250 | Frequency of unknown/not stated (etc) |
| Test Result Type | 12608.5 | Test result type | Sample tracking - GMC QC Test | If: 'Tumour content'; 'Cellularity'; 'Percent necrosis' | NA | |
| Test Result Value | 12610.1 | Test result value | Sample tracking - GMC QC Test | If Test result type = 'Tumour content' 'Cellularity'; 'Percent necrosis' | If 'Tumour content': Low; Medium; High. If 'Cellularity': VeryLow; Low; Medium; High; VeryHigh. If 'Percent necrosis': 0-100 (ranges allowed) | 'Percent necrosis': Frequency of unknown (etc) |
| Morphology (ICD) [Core] | 14871.2 | Icd Code | Diagnosis/2 | At least one required | Format: 4 - 6 alphanumeric digits | |
| Morphology (SnomedRT) [Core] | 31243.1 | Snomed Rt Code | Diagnosis/2 | | Code validity check against online database. ≠ M09350 (morphological description only); ≠ M09450 (no evidence of malignancy) | |
| Morphology (SnomedCT) [Core] | 31244.1 | Snomed Ct Code | Diagnosis/2 | | Code validity check against online database. ≠ M09350 (morphological description only); ≠ M09450 (no evidence of malignancy) | |
| Topography (ICD) [Core] | 31228.1 | Icd Code | Diagnosis | At least one required | Format: 4 - 6 alphanumeric digits | Produce listing where Topograph (ICD) ≠ Diagnosis (ICD), per GMC |
| Topography (SnomedRT) [Core] | 31227.1 | Snomed Rt Code | Diagnosis | | Code validity check against online database. ≠ M09350 (morphological description only); ≠ M09450 (no evidence of malignancy) | |

4 Definitions & Abbreviations

| Abbreviation / Term | Description |
|---------------------|--|
| COSD | Cancer Outcomes and Services Dataset |
| DC | Data Clarification |
| DQ | Data Quality |
| GEL | Genomics England |
| GMC | Genomics Medicine Centre |
| GELDAM | Genomics England Data Acquisition Management Systems |
| HES | Hospital Event Statistics |
| PHE | Public Health England |

5 Related Documents, References & Procedures

Include here relevant references, documents and procedures which directly link to the procedural activities described.

Do not include general references which are applicable project-wide, as these are referenced in key project documents.

6 Appendices

6.1 Document Version History

| Version | Author | Date | Description |
|--------------------|--|-----------|---------------------------|
| V0.1 – V0.5 | Rob Jackson, Clinical Data Manager | 01-NOV-16 | Draft versions for review |
| V1.0 | Amanda O' Neill, Director of Clinical Data | 15-NOV-16 | First approved version |