TECHNICAL SPECIFICATION

ISO/TS 20428

First edition
2017-05

Health informatics — Data elements and their metadata for describing structured clinical genomic sequence information in electronic health records

Informatique de santé — Éléments de données et leurs métadonnées pour décrire l'information structurée de la séquence génomique clinique dans les dossiers de santé électroniques

Reference number
ISO/TS 20428:2017(E)

ISO copyright office
Ch. de Blandonnet 8 • CP 401
CH-1214 Vernier, Geneva, Switzerland
Tel. +41 22 749 01 11
Fax +41 22 749 09 47
copyright@iso.org
www.iso.org

Contents

Page

Foreword

ISO (the International Organization for Standardization) is a worldwide federation of national standards bodies (ISO member bodies). The work of preparing International Standards is normally carried out through ISO technical committees. Each member body interested in a subject for which a technical committee has been established has the right to be represented on that committee. International organizations, governmental and non-governmental, in liaison with ISO, also take part in the work. ISO collaborates closely with the International Electrotechnical Commission (IEC) on all matters of electrotechnical standardization.

The procedures used to develop this document and those intended for its further maintenance are described in the ISO/IEC Directives, Part 1. In particular the different approval criteria needed for the different types of ISO documents should be noted. This document was drafted in accordance with the editorial rules of the ISO/IEC Directives, Part 2 (see www.iso.org/directives).

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. ISO shall not be held responsible for identifying any or all such patent rights. Details of any patent rights identified during the development of the document will be in the Introduction and/or on the ISO list of patent declarations received (see www.iso.org/patents).

Any trade name used in this document is information given for the convenience of users and does not constitute an endorsement.

For an explanation on the voluntary nature of standards, the meaning of ISO specific terms and expressions related to conformity assessment, as well as information about ISO's adherence to the World Trade Organization (WTO) principles in the Technical Barriers to Trade (TBT) see the following URL: www.iso.org/iso/foreword.html.

The committee responsible for this document is ISO/TC 215, *Health informatics*.

Introduction

Based on the rapid advancement of sequencing technologies, clinical sequencing has been highlighted as one of methods to realize personalized medicine and precision medicine. There are lots of sequencing data in the public domain with clinical information[1]. In addition, genome-scale clinical sequencing is being adopted broadly in medical practice[2]. Many hospitals have started to sequence patients' whole genome, whole exome, or targeted genes using the next generation sequencing technologies. These genomic data obtained by next generation sequencing technologies can be used for both clinical purposes to diagnose patients and choose the right medications and research purposes. Therefore, the management of genomic and clinical data is increasingly highlighted to precision medicine, clinical trial and translational research[3].

However, until now, there is no international standard for representing clinical sequencing results with a structured format for electronic health records, in consequence, the necessary genomic test results are not efficiently delivered to the clinicians. There are a few related standards for modelling genetic testing results (i.e. ISO 25720 and several HL7 documents from HL7 clinical genomics working group). However, these standards or drafts are mainly focused on the traditional genetic testing results for a single gene test. Based on the rapid development and adoption of next generation sequencing techniques which can detect diverse genetic variants in genome level, there is, therefore still a need to develop a standard to present clinical sequencing data in such a way they become useful for clinicians[4].

To implement a structured clinical sequencing report in electronic health records, all necessary data fields should be defined and the metadata for each chosen field should be defined. For example, it needs to be determined which vocabulary, in particular gene descriptions and/or disease codes, can be applied in particular fields. In ISO TC 215, GSVML (Genomic Sequence Variation Markup Language) was proposed for interoperability of genomic variants, especially for single nucleotide polymorphism (SNP) data[5]. HL7 is also developing a domain analysis model for genomics using HL7 version 3[6] and fast healthcare interoperability resources (FHIR)[7]. Recently, to facilitate genomic information, SMART on FHIR Genomics has been developed[8],[9]. The Clinical Data Interchange Standard Consortium (CDISC) published a study data tabulation model implementation guide: pharmacogenomics/genetics[10]. Several other international organizations such as Global Alliance for Genomics and Health (GA4GH), Actionable Genome Consortium, and Displaying and Integrating Genetic Information Through the EHR (DIGITizE) of Institute of Medicine in US, tried to develop the similar standards. The working group of the American College of Medical Genetics and Genomics Laboratory Quality Assurance Committee published the ACMG clinical laboratory standards for next-generation sequencing[11]. In addition, web-based tools become available that link genotypic information to phenotypic information, and exchanging information and using it in personalized medicine can be very helpful[12].

In this document, to enable the standard use of patient genomic data from clinical sequencing for healthcare purposes as well as for clinical trials and research, the metadata for a clinical sequencing report for electronic health records will be developed. It further explains how and where particular appropriate terminological systems that describe the genomes and/or diseases can be applied in these fields. By defining the necessary fields with structured format based on coded data that adhere themselves to terminological principles such as concept representation and governance, this document can help implement clinical decision support service.

Health informatics — Data elements and their metadata for describing structured clinical genomic sequence information in electronic health records

1 Scope

The document defines the data elements and their necessary metadata to implement a structured clinical genomic sequencing report and their metadata in electronic health records particularly focusing on the genomic data generated by next generation sequencing technology.

This document

— defines the composition of a structured clinical sequencing report (see Clause 5),

— defines the required data fields and their metadata for a structured clinical sequencing report (see Clause 6),

— defines the optional data (see Clause 7),

— covers the DNA-level variation from human samples using whole genome sequencing, whole exome sequencing, and targeted sequencing (disease-targeted gene panels) by next generation sequencing technologies. Though whole transcriptome sequencing and other technologies are important to provide better patient care and enable precision medicine, this document only deals with DNA-level changes,

— covers mainly clinical applications and clinical research such as clinical trials and translational research which uses clinical data. However, the necessary steps such as de-identification or consent from patient should be applied. The basic research and other scientific areas are outside the scope of this document,

— does not cover the other biological species, i.e. genomes of viruses and microbes, and

— does not cover the Sanger sequencing methods.

2 Normative references

There are no normative references in this document.

3 Terms and definitions

For the purposes of this document, the following terms and definitions apply.

ISO and IEC maintain terminological databases for use in standardization at the following addresses:

— ISO Online browsing platform: available at http://www.iso.org/obp

— IEC Electropedia: available at http://www.electropedia.org/

3.1
allele
one of several alternate forms of a gene which occur at the same locus on homologous chromosomes and which become separated during meiosis and can be recombined following fusion of gametes

[SOURCE: ISO 16577:2016, 3.6]

3.2
benign
alterations with very strong evidence against pathogenicity

3.3
biomaterial
materials taken from the human body such as tissue, blood, plasma or urine

3.4
chromosome
structure that comprises discrete packages of DNA and proteins that carries genetic information which condense to form characteristically shaped bodies during nuclear division

[SOURCE: ISO 19238:2014, 2.7]

3.5
clinical sequencing
next generation sequencing or later sequencing technologies with human samples for clinical practice and clinical trials

3.6
ClinVar
freely accessible, public archive of reports of the relationships among human variations and phenotypes, with supporting evidence

Note 1 to entry: http://www.ncbi.nlm.nih.gov/clinvar/.

3.7
copy number variation
CNV
variation in the number of copies of one or more sections of the DNA

3.8
Catalogue of Somatic Mutations in Cancer
COSMIC
online database of somatically acquired mutations found in human cancer

Note 1 to entry: http://cancer.sanger.ac.uk/cosmic.

3.9
dbSNP
database of SNPs provided by the US National Center for Biotechnology Information (NCBI)

Note 1 to entry: https://www.ncbi.nlm.nih.gov/SNP/.

3.10
deletion
mutation in which a part of a chromosome or a sequence of DNA is lost during DNA replication

3.11
deoxyribonucleic acid
DNA
molecule that encodes genetic information in the nucleus of cells

[SOURCE: ISO 25720:2009, 4.7]

3.12
DNA sequencing
determining the order of nucleotide bases (adenine, guanine, cytosine and thymine) in a molecule of DNA

Note 1 to entry: Sequence is generally described from the 5' end.

[SOURCE: ISO/TS 17822-1:2014, 3.20]

3.13
exome
part of the genome formed by exons

3.14
gene
basic unit of hereditary material that encodes and controls the expression of a protein or protein subunit

[SOURCE: ISO 11238:2012, 2.1.16]

3.15
gene panel
technique for sequencing the targeted genes in a genome

3.16
genomic medicine
medical discipline that involves using genomic information about an individual as part of their clinical care (e.g. for diagnostic or therapeutic decision-making) and the health outcomes and policy implications of that clinical use

3.17
germline
series of germ cells each descended or developed from earlier cells in the series, regarded as continuing through successive generations of an organism

3.18
indel
insertion (3.19) or/and *deletion* (3.10)

3.19
insertion
addition of one or more nucleotide base pairs into a DNA sequence

3.20
inversion
chromosome rearrangement in which a segment of a chromosome is reversed end to end

3.21
large indel
insertion or deletion up to ~1 kb

3.22
likely benign
alterations with strong evidence against pathogenicity

Note 1 to entry: Targeted testing of at-risk family members not recommended.

3.23
likely pathogenic
alterations with strong evidence in favor of pathogenicity

3.24
pathogenic
characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention

3.25
prenatal/fetal
biomaterial sample of fetuses before birth

Note 1 to entry: Prenatal/fetal DNA sequencing: Reading the DNA of foetuses to diagnose Mendelian disease of unborn child.

3.26
sequence read
read
fragmented nucleotide sequences which are used to reconstruct the original sequence for next generation sequencing technologies

3.27
read type
type of run in the sequencing instrument

Note 1 to entry: It can be either single-end or paired-end.

Note 2 to entry: Single-end: Single read runs the sequencing instrument reads from one end of a fragment to the other end.

Note 3 to entry: Paired-end: Paired end runs read from one end to the other end, and then start another round of reading from the opposite end.

3.28
reference sequence
digital nucleic acid sequence database, assembled by scientists as a representative example of human genome

3.29
ribonucleic acid
RNA
polymer of ribonucleotides occurring in a double-stranded or single-stranded form

[SOURCE: ISO 22174:2005, 3.1.3]

3.30
sequence variation
DNA sequence variation
variation
differences of DNA sequence among individuals in a population

Note 1 to entry: Variation implies *CNV* (3.7), *deletion* (3.10), *insertion* (3.19), *indel* (3.18), *small indel* (3.32), *large indel* (3.20), and *SNP* (3.31).

[SOURCE: ISO 25720:2009, 4.8]

3.31
single nucleotide polymorphism
SNP
single nucleotide variation in a genetic sequence that occurs at appreciable frequency in the population

Note 1 to entry: Pronounced “snip”.

[SOURCE: ISO 25720:2009, 4.23]

3.32
small indel
insertion or deletion of 2 ~100 nucleotides

3.33
somatic cell
cells of the body in contrast to the germ line cells

3.34
biological specimen
biospecimen
specimen
sample of tissue, body fluid, food, or other substance that is collected or acquired to support the assessment, diagnosis, treatment, mitigation or prevention of a disease, disorder or abnormal physical state, or its symptoms

3.35
subject of care
any person who uses, or is a potential user of, a health care service

[SOURCE: ISO/TS 22220:2011, 3.2]

3.36
target capture
method to capture genomic regions of interest from a DNA sample prior to sequencing

3.37
uncertain significance
alterations with limited and/or conflicting evidence regarding pathogenicity

3.38
whole exome sequencing
WES
technique for sequencing all the protein-coding genes in a genome

3.39
whole genome sequencing
WGS
technique that determines the complete DNA sequence of an organism's genome at a single time

4 Abbreviated terms

This list of abbreviated terms includes all abbreviations used in this document.

ACMG	the American College of Medical Genetics and Genomics
COSMIC	the Catalogue of Somatic Mutations in Cancer
CPIC	the Clinical Pharmacogenetics Implementation Consortium
EBI	the European Bioinformatics Institute
FHIR	Fast Healthcare Interoperability Resources
HGNC	the HUGO Gene Nomenclature Committee
HGVS	the Human Genome Variation Society
HUGO	the Human Genome Organization
IARC	International Agency for Research on Cancer
LOINC	Logical Observation Identifiers Names and Codes

NCBI	National Center for Biotechnology Information
NCCN	National Comprehensive Cancer Network
NGS	Next Generation Sequencing
SNP	Single Nucleotide Polymorphism
SPREC	Standard Preanalytical Code
WES	Whole Exome Sequencing
WGS	Whole Genome Sequencing

5 Use case scenario

The abstracted use case for generating a clinical genomic sequencing report is demonstrated in Figure 1. At first, the clinician will place a clinical sequencing order using the electronic health records system (step 1 in Figure 1). After the order, a responsible department will request DNA sequencing to the sequencing facility (step 2). This sequencing facility can be located inside of the hospital or be an independent sequencing facility outside the hospital (step 3). When confirming the order, the sequencing facility will request a sample from the patient (step 4). The hospital will collect a sample from the patient (steps 5 and 6). The pre-collected samples, i.e. biobank sample, the samples acquired by a previous laboratory or pathology orders, can be used as well. The biomaterial from the patients will be delivered to the sequencing facility (step 7). After receipt the sequencing facility will perform a sequencing analysis (step 8) and generate the report (step 9). This report will be sent to the requested hospital (step 10), and the report will be updated in the electronic health record system (step 11). The ordering clinician will be notified of the completion of the sequencing order (step 12). Finally, the ordering clinician will make a diagnosis or give a proper treatment (step 13). A patient can have a copy of final report.

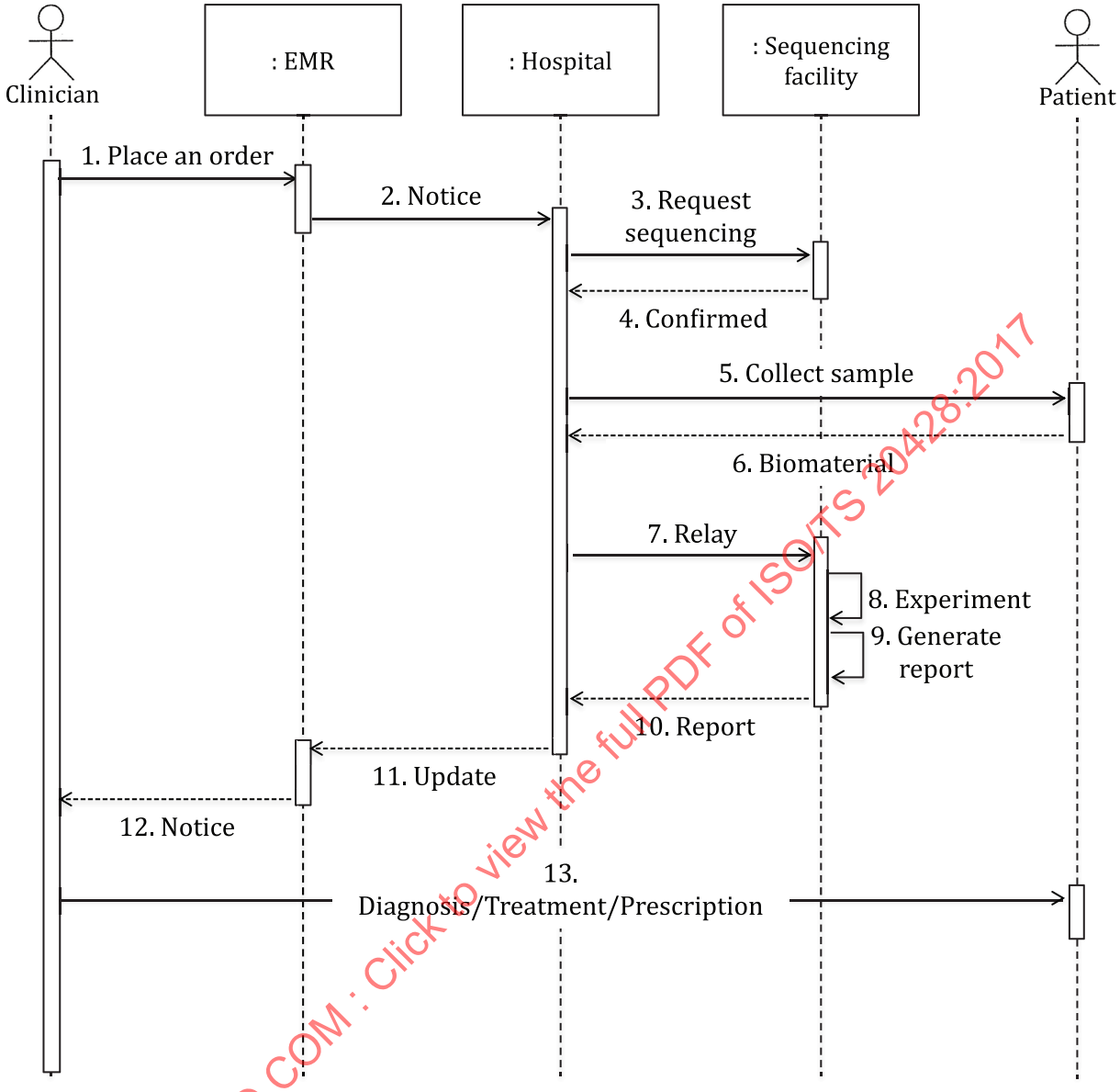

Figure 1 — Abstracted flow of generating clinical genomic sequencing report

The possible use cases for the clinical sequencing orders are well described in "HL7 VERSION 3 Domain Analysis Model: Clinical Sequencing"[6].

6 Composition of a clinical sequencing report

6.1 General

The structured clinical sequencing report may mainly consist of two parts: the summary part and the detailed contents part as in Figure 2. The summary part can include the subset of required fields to help clinicians quickly overview the most important findings concisely[13],[14]. The detailed content part should contain all required fields and the selected optional fields.

This document only defines the data elements and their metadata for the structured clinical sequencing report in electronic health records. Therefore, its layout can be designed based on the institutional decision if all elements are included as in this document.

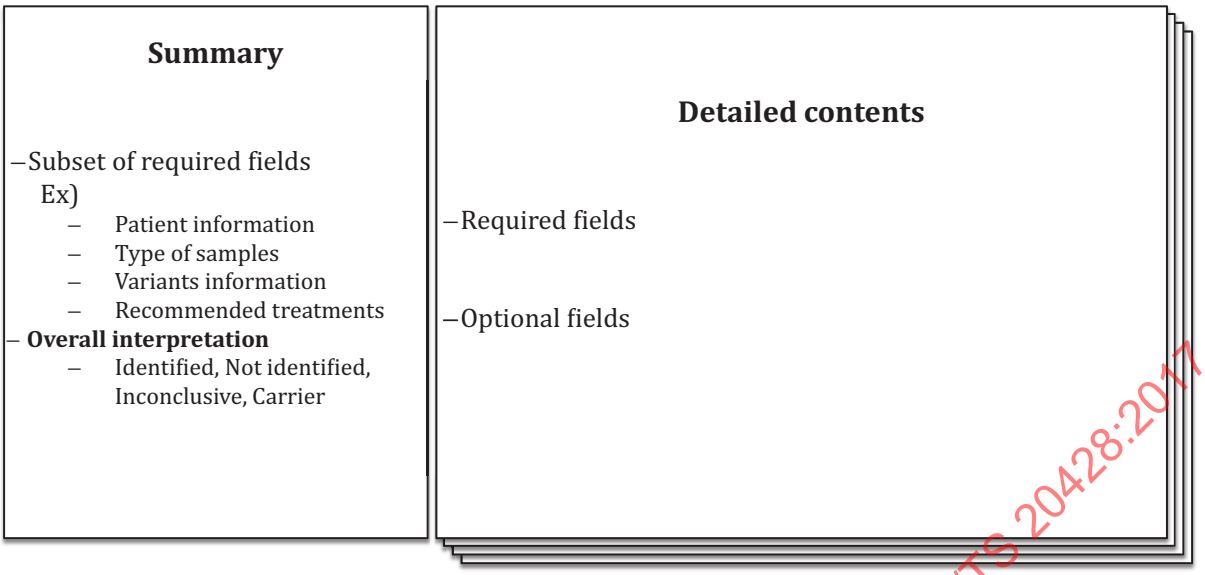

Figure 2 — Composition of a clinical sequencing report

6.2 Overall interpretation in summary

The summary part should report the overall interpretation of a genomic test with a succinct description: identified, not identified inconclusive, or carrier. Table 1 summarizes each interpretation. "Identified" represents a detection of a variant that explains a patient's condition. "Not identified" means no variants identified of likely relevance to the diagnostic indication. "Inconclusive" is a clear explanation of the patient's condition was not found either due to only variants of unknown significance being identified or due to only a single heterozygous variant identified for a recessive condition. "Carrier" represents the identification of variants of recessive carrier screening tests[11].

Table 1 — Overall interpretation in the summary part

Interpretation	Remarks
Identified	detection of a variant that explains a patient's condition
Not identified	no variant identified of likely relevance to the diagnostic indication
Inconclusive	a clear explanation of the patient's condition was not found
Carrier	identification of variants of recessive carrier screening tests

6.3 Detailed contents

The detailed contents consist of two parts such as the required and optional fields. The detailed content part should contain all required fields and the selected optional fields. HL7 Implementation Guide for CDA Release 2 Genetic Testing Report can be a good example. However, HL7 Genetic Testing Report only focuses on the single gene test and does not cover next generation sequencing technology.

The required fields mainly focus on helping clinicians by providing the necessary genomic information, interpretation results, and the related treatments. They include all necessary fields for clinical practice. The optional fields give more detailed information to clinicians. They also can facilitate translational research with the necessary steps such as de-identification or consent from the patient. In Reference [13], the data fields for molecular genetic report template were categorized into "required," "optional," "possible," and "not necessary" based on the survey. The data fields in this document were re-categorized to implement a structured clinical sequencing report by reviewing the existing free-text format sequencing report.

7 Fields and their nomenclature of required data

7.1 General

The required fields are chosen for clinical practice. The information which can be only described in the clinical sequencing report is included in the required fields to minimize the length of clinical sequencing report. The other relevant information can be included in the optional fields or other clinical reports in the electronic health records. The summary of data elements and their metadata are shown in Table 2.

Table 2 — Data elements and their metadata for required fields

Data elements			Metadata (Primary)
Clinical sequencing orders	Clinical sequencing order code	Order code	LOINC
		Information on sequencing order	TEXT
	Date and time	Order date	ISO 8601
		Specimen collection date	
		Order received date	
		Report date	
		Addendum creation date	
	Specimen information		ISO/TS 22220:2011
Information on subject of care	Identifiers		ISO/TS 22220:2011
	Name		
	Birth date		ISO 8601
	Sex		ISO/TS 22220:2011
	Ethnicity		HL7 v3 Code System Race
Information of legally authorized person ordering clinical sequencing			ISO/TS 27527:2010
Performing laboratory	Basic information		TEXT
	Information of report generator		TEXT
	Information of legally confirmed person on sequencing report		ISO/TS 27527:2010
Associated diseases and phenotypes			ICD
Biomaterial information	Type of sample		SPREC
	Genomic source class in biomaterial		LOINC
	Conditions of specimen		TEXT
Genetic variations	Gene symbols and names		HGNC
	Sequence variation information	Notation	HGVS
		Effects of variants	TEXT
		Sequence variant ID	Database unique ID
Classification of variants	Pathogeny		ENUM (“Pathogenic”, “Likely pathogenic”, “Unknown significance”, “Likely benign”, “Benign”)[a]
	Clinical relevance		ENUM (“Identified,” “Likely identified”, “Uncertain”, “Not identified”)

[a] ENUM represents the contents should be chosen among the given category.

Table 2 *(continued)*

Data elements		Metadata (Primary)
Recommended treatment	Medication	ISO 11615
	Clinical trial information	Clinical trial ID
	Known protocols related to a variant	TEXT
	Other recommendation	TEXT

[a] ENUM represents the contents should be chosen among the given category.

7.2 Clinical sequencing orders

7.2.1 General

When a clinician orders a clinical genomic sequencing, the order code and the required date information should be given.

7.2.2 Clinical sequencing order code

7.2.2.1 Order code

The relevant clinical sequencing orders should be represented by LOINC (Logical Observation Identifiers Names and Codes) with used LOINC coding system version. Unfortunately, there are not LOINC codes for those sequencing orders, but only molecular genetic code until now. In the meantime, other international, national, or institutional coding system can be alternatively used, for example, CPT (Current Procedure Terminology) in US has the codes for sequencing.

7.2.2.2 Information on sequencing order

If order code cannot fully describe the purpose of the clinical sequencing, the detailed description of the sequencing can be added as free text, for example, the sequencing panel name with gene information.

7.2.3 Date and time

For a clinical sequencing report, diverse date and time should be reported due to time delay. All date and time in the report should be represented by ISO 8601.

7.2.3.1 Order date

Order date is the date which clinician ordered the necessary clinical genomic sequencing.

7.2.3.2 Order received date

Order received date is the date that the performing laboratory received and confirmed the clinical genomic sequencing order.

7.2.3.3 Specimen collection date

Specimen collection date indicates the date when specimen taken from patient or tissue.

7.2.3.4 Report date

Report date is the date which the performing laboratory generates the sequencing report.

7.2.3.5 Addendum creation date

Addendum creation date is the date that the performing laboratory creates the addendum of the previous report using up-to-date information. The analysis pipelines or referred database is updated due to technology advancement. The reference sequence is regularly updated. In addition, it should be updated due to legal aspects. Therefore, the performing laboratory should create the addendum of the existing sequencing report based on the clinician's request or law enforcement.

7.2.4 Specimen information

The specimen information can be represented by subject of care identifier type code of ISO 22220:2011.

EXAMPLE 13-S-048435_A1 - Pathology Number: ISO 22220:2011 (SOC identifier designation: 13-S-048435_A1, SOC identifier geographic area: 1 (local), SOC identifier issuer: AMC (ABC Medical Center), 02 (speciality number - pathology).

7.3 Information on subject of care

7.3.1 General

Information on subject of care whose specimen was sequenced should be represented by ISO 22220:2011.

EXAMPLE 12345678 - ISO 22220:2011 (SOC identifier designation: 12345678, SOC identifier geographic area: 1 (local), SOC identifier issuer: AMC (ABC Medical Center), 01 (unique identifier for issuer).

7.3.2 Subject of care identifiers

The unique identifiers for subjects of care should be included.

7.3.3 Subject of care name

The subject of care name should be given.

7.3.4 Subject of care birth date

The subject of care's birth date should be given to calculate the patient's age. Birth date should be represented by ISO 8601.

7.3.5 Subject of care sex

The subject of care's sex should be represented by ISO 22220:2011.

7.3.6 Subject of care ethnicity

The ethnicity of the subject of care should be notified to represent his or her genetic origin. The ethnicity information should be represented by HL7 v3 Code System Race (https://www.hl7.org/fhir/v3/Race/index.html). Alternatively, if there are national standards, those coding systems can be used, for example, US FDA guidance for Industry – Collection of Race and Ethnicity Data in Clinical Trials[15].

7.4 Information on legally authorized person ordering clinical sequencing

7.4.1 General

Information on legally authorized person who ordered clinical sequencing might be represented by ISO/TS 26527:2010.

The detailed items of this field such as the name of ordering physician, his/her medical speciality, or contact numbers, can be chosen by the implementing hospitals or laboratories.

7.5 Performing laboratory

7.5.1 General

The laboratory that performs the sequencing should be given.

7.5.2 Basic information on performing laboratory

The name of performing laboratory and contact points such as phone numbers or emails should be notified. This information can be given in free text until the relevant standard will be published.

7.5.3 Information on report generator

Information on subject of provider who generated a report of sequencing results should be reported. It can be represented by ISO/TS 26527:2010 or by free text.

7.5.4 Information of legally confirmed person on sequencing report

Information of legally confirmed physician might be given by ISO/TS 26527:2010.

7.6 Associated diseases and phenotypes

If possible, associated diseases and phenotypes should be included using ICD codes or SNOMED-CT code. Other phenotypes that are not classified by ICD or SNOMED-CT can be represented by Human Phenotype Ontology (http://human-phenotype-ontology.github.io/).

Since even single variant can have multiple associated phenotypes including diseases, all possible associated diseases should be listed based on the confidence level.

In this field, only diseases or phenotypes those are associated with the found variants should be indicated. The previously known diseases of subject of care, which are not associated with the variants, should be excluded in this field.

7.7 Biomaterial information

7.7.1 General

The information on the specimen from patient should be described.

7.7.2 Types of sample

Types of samples can be represented by Standard Preanalytical Code (SPREC) V2.0 of International Society for Biological and Environmental Repositories[16].

EXAMPLE BLD (Blood), BUF (buffy coated), non-blood tissue (CEN), semen (SEM).

7.7.3 Genomic source class in biomaterial

The genomic source of patient's sample should be categorized as follows: 1) germline, 2) somatic, 3) prenatal/fetal, 4) likely germline, 5) likely somatic, 6) likely prenatal/fetal, or 7) unknown genomic origin. This category is based on LOINCS answer list LL378-1.

7.7.4 Conditions of specimen that may limit adequacy of testing

The specific conditions that can affect or limit the adequacy of genomic sequencing should be described in text format.

EXAMPLE Tumor purity information should be given in percentage.

7.8 Genetic variations

7.8.1 General

All found variants should be listed in the report according to their relevance to the patient's indication for testing. However, in the required field, the variants that have the associated treatments should be listed. The other variants can be listed in the optional fields.

As in 7.9, classification of variants can be categorized based on the founding variant has a proper treatment. In the clinical setting, the variants without proper treatment have no meaning to the clinicians. Therefore, only variants with treatments should be included in the required field.

The variants information should include the following information.

7.8.2 Gene symbols and names

The gene symbol and name which contains the founding variants should be represented by HGNC (HUGO (Human Genome Organization) Gene Nomenclature Committee). The HGNC approved gene symbol and HGNC ID can be used. The symbol is case-insensitive.

EXAMPLE HGNC approved symbol: BRCA1, HGNC ID: HGND:1100.

HGNC is a non-profit body which is jointly funded by the US National Human Genome Research Institute (NHGRI) and the Welcome Trust (UK). They operate under the auspices of HUGO (Human Genome Organization).

7.8.3 Sequence variation information

7.8.3.1 General

The sequence variant should be represented by HGVS (Human Genome Variation Society, http://www.hgvs.org/) nomenclature. Amino acid changes can be followed to give more information.

EXAMPLE c.76A>T_ p.Asn26Tyr.

7.8.3.2 HGVS nomenclature

The detailed explanation on description of sequence changes in DNA level can be found at http://varnomen.hgvs.org/recommendations/DNA/.

EXAMPLE The examples of diverse sequence variants can be presented as follows:

Substitutions: In HGVS, ">" indicates a substitution at DNA level (i.e. c.76A>T).

Deletions: In HGVS, "del" indicates a deletion (i.e. c.76delA).

Duplications: In HGVS, "dup" indicates a duplication (i.e. c.76dupA).

Insertion: In HGVS, "ins" indicates an insertion (i.e. c.76_77insG).

Insertion/deletion: In HGVS, "delins" indicates an indel (c.112_117delinsTG). Indels are described as a deletion followed by an insertion.

Inversions: In HGVS, 'inv' indicates an inversion (i.e. c.76_83inv).

Conversions: In HGVS, 'con' indicates a conversion. The example "g.123_678conNG_012232.1:g.9456_10011" describes a gene conversion replacing nucleotides 123 to 678 of the reference genomic sequence with nucleotides 9456 to 10011 from the sequence as present in GenBank file NG_012232.1.

Translocations: In HGVS, translocations are described at the molecular level using the format "t(X;4)(p21.2;q35)", followed by the usual numbering, indicating the position translocation breakpoint. t(X;4)(p21.2;q35)(c.857+101_857+102) denotes a translocation breakpoint in the intron between coding DNA nucleotides 857+101 and 857+102, joining chromosome bands Xp21.2 and 4q34.

7.8.3.3 Effects of variants

Type of variants can be explained using free texts.

EXAMPLE Substitution, Deletion, Duplication, Insertion, Inversion, and Conversion.

Additionally, the effects of mutation on protein function such as missense, nonsense, and silent can be given with parentheses.

EXAMPLE Substitution (missense).

7.8.3.4 Sequence variant ID

All known variants can be reported using other unique IDs.

If sequence variant has dbSNP ID (http://www.ncbi.nlm.nih.gov/projects/SNP/), the variant can be represented by dbSNP ID. The prefix "rs" represents dbSNP ID (i.e. rs10000). dbSNP is maintained by NCBI.

ClinVar is a database for the relationships among human variations and phenotypes, with supporting evidence, maintained by NCBI (http://www.ncbi.nlm.nih.gov/clinvar/). If ClinVar ID is used, the identifier should be prefixed with "ClinVar:" (i.e. ClinVar:17661) as in HL7 FHIR Standard Profile for Genetics.

If a variant is a somatic variant, COSMIC (Catalogue of somatic mutations in cancer) identifier can be used with other identifiers (http://cancer.sanger.ac.uk/cosmic). COSMIC id should be prefixed with "COSM" (i.e. COSM12979) as in HL7 FHIR Standard Profile for Genetics. COSMIC is a database to store and display somatic mutation information and related details and contains information relating to human cancers. COSMIC is maintained by the Wellcome Trust Sanger Institute.

7.9 Classification of variants

7.9.1 General

Currently, there are two different approaches to classify genetic variants: one focuses on the pathogeny and the other focuses on clinical relevance based on the existence of possible medication. These two classifications have different meanings, the institution can temporally choose which one will be used or make a new classification by combining two existing classifications.

7.9.2 Classification of variants based on the pathogeny

This classification follows the ACMG recommendations for standards for interpretation and reporting of sequence variations: Revision 2015[17], IARC 5-class system[18], or College of American Pathologists guideline[19]: Pathogenic, likely pathogenic, unknown significance, likely benign, and benign.

Pathogenic: Segregates with disease in >3 unrelated cases with control data. The variants are located in a highly-conserved region. Functional studies or other evidences suggest deleterious effect on gene expression.

Likely pathogenic: The variants are reported in a few case studies with or without control data. Functional studies or other evidence suggests deleterious effect on gene expression.

Unknown significance: Nothing has been reported regarding this variant, or reported information is incomplete and/or contradictory.

Likely benign: Variants are reported in few cases, with little or without control data. May be non-conserved and/or predicted to be well tolerated. Frequency is higher than expected in the general population based on disease prevalence and penetrance. And the variants are identified as functionally normal.

Benign: Reported at high frequency in control data/does not segregate with disease. Variant may be non-conserved and/or predicted to be well tolerated. And the variants are identified as functionally normal.

7.9.3 Classification of variants based on clinical relevance

In a clinical setting, alternative classification is necessary to provide a proper treatment for patient[20]. In this document, we revised the classification: Identified, likely identified, uncertain, and not identified.

Identified: The clinically relevant (actionable) variants are detected.

Likely identified: The variants with likely clinical relevance are detected.

Uncertain: The variants with uncertain clinical relevance are detected.

Not identified: No clinically relevant variants are detected.

7.10 Recommended treatment

7.10.1 General

The recommend treatment such as medication or clinical trials can be reported to help clinicians.

7.10.2 Classification of variants based on clinical relevance

The associated medication can be represented by MPID (Medicinal Product IDentifier) or IMPID (Investigational MPID) of ISO 11615 IDMP (Identification of Medicinal Product) Standard. The Anatomical Therapeutic Chemical (ATC) classification maintained by the WHO Collaborating Centre for Drug Statistics Methodology, can be alternatively used to represent the classification of active ingredients of drugs. International Nonproprietary Names (INN) facilitates the identification of pharmaceutical substances or active pharmaceutical ingredients. Each INN is a unique name that is globally recognized and is public property. If INN exists, INN can be alternatively used as drug name.

However, it will be most directly beneficial to individual patients for the recommended target of therapy to be represented using the same coding scheme that has already used within the particular country or electronic health record system, so that the treatment recommendation and the actual prescription or course of chemotherapy are semantically aligned within the record of a single patient, who is likely to be treated in a local environment rather than globally. Therefore, if there is a national or local standard, it can be used temporarily. The examples are RxNorm, which is maintained by the US. National Library of Medicine, or KD Code (Korea Drug Code), which is maintained by the Korea Health Insurance Review and Assessment Service.

7.10.3 Clinical trial information

The clinical trials which test the founding variants-targeted drugs can be given to help clinicians. The clinical trial information should be represented by ClinicalTrials.gov ID or EudraCT number. Other domestic registry ID can be alternatively used.

ClinicalTrials.gov is a registry of federally and privately supported clinical trials conducted in the United States and around the world. If there are the related clinical trials to detected variants, the ClinicalTrials.gov ID can be given. In addition, target condition, intervention, phase, and study results can be attached.

Example NCT00844506.

EudraCT (https://eudract.ema.europa.eu/) is a database of all clinical trials which commenced in Europe. EudraCT number also can be used to represent the related clinical trials.

7.10.4 Known protocols related to a variant

If the found variants are classified as clinical actionable variants, the established clinical guidelines should be notified.

EXAMPLES CPIC (Clinical Pharmacogenetics Implementation Consortium) guideline: https://www.pharmgkb.org/view/dosing-guidelines.do?source=CPIC.[21]

NCCN (National Comprehensive Cancer Network) guideline: http://www.nccn.org/professionals/physician_gls/f_guidelines.asp.

7.10.5 Other recommendation

Other recommendation can be included, for example, other laboratory tests including another sequencing order or testing relatives.

7.11 Addendum

If there is an addendum, it should be described in the required field. The reason of addendum creation can be included.

8 Fields and their nomenclature of optional data

8.1 General

The data elements for optional fields in the structured clinical sequencing report focused on mainly the support of clinical decision by giving more detailed information. These fields can also be applied to clinical trial and translational research.

The fields that are not listed in this document can be also used based on institutional decision.

Table 3 — Data elements and their metadata for optional fields

Data elements			Metadata (Primary)
Medical history			ICD
Family history/Pedigree information			HL7 v3 IG: Family History/Pedigree Interoperability,
Reference genome version			Genome Reference Consortium Human Genome release ID
Racial genome information			TEXT
Genetic variation	Gene symbols and names		HGNC
	Sequence variation information	Notation	HGVS
		Effects of variants	TEXT
		Sequence variant ID	Database unique ID
	HGVS version		HGVS version number

[a] ENUM represents the contents should be chosen among the given category.

Table 3 *(continued)*

Data elements			Metadata (Primary)
Detailed sequencing information	Clinical sequencing date		ISO 8601
	Quality control metrics		NUMERIC
	Base calling information	Read depth	NUMERIC
		Reference allelic depth	
		Alternative allelic depth	
		Allele frequency	
		Genotype	
	Sequencing platform information	Type of sequencers	TEXT
		Library preparation methods	
		Target capture methods	
		Read type	ENUM ("single-end", "paired-end")
		Read length	TEXT
	Analysis platform information	Alignment tools	TEXT
		Variant calling tools	
		Other tools	
		Chromosome coordination system	ENUM ("zero-based", "one-based", "zero-based, half-open")
		Annotation tools and databases	TEXT
References			TEXT

[a] ENUM represents the contents should be chosen among the given category.

8.2 Medical history

The medical history of subject of care can be reported. When reporting medical history, the relevant standard terminology should be used, for example, ICD for disease names or IDMP for medication.

8.3 Family history/Pedigree information

All family history or pedigree information should be represented by HL7 v3 Implementation Guide: Family History/Pedigree Interoperability, Release 1.

8.4 Reference genome version

Reference sequences are the baseline from which variation is reported. If different reference sequences are used, the variant calls are also different. The reference sequence should be represented by Genome Reference Consortium Human Genome release ID or Locus Reference Genomic ID.

Reference sequences should be represented by Genome Reference Consortium Human Genome release ID (http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/human/). If there is an update, the patch number should be appended.

EXAMPLE GRCh38 or GRCh37.p13 (GRCh37 Patch Release 13)

LRG (Locus Reference Genomic) ID (http://www.lrg-sequence.org/), which is maintained by EBI, can be used as well.

EXAMPLE LRG_1.

8.5 Racial genomic information

When racial genomic information was reported, the reference data set for racial information should be notified using HapMap (http://hapmap.ncbi.nlm.nih.gov/), 1000 Genomes (http://www.1000genomes.org/), Exome Aggregation Consortium (ExAC) data set (http://exac.broadinstitute.org/), Genome Aggregation Database (gnomAD) (http://gnomad.broadinstitute.org/), or other sequencing database which can give racial information.

8.6 Genetic variation

The remaining variants, which are not reported in the required part, can be listed in the optional part. The variants should be reported followed by 7.7. However, in optional fields, HGVS version should be notified.

EXAMPLE HGVS Version 15.11.

8.7 Detailed sequencing information

Data that is automatically generated through a sequencing machine should be given for research purpose. Until now, there is no standard for this purpose. The notation used in conventional bioinformatics fields can be used alternatively.

8.7.1 Clinical sequencing date

Clinical sequencing date is the date which the performing laboratory generates the sequencing results using the received specimen. Usually, after receiving samples, the laboratory gathers the enough samples to run the sequencer to optimize the efficiency of sequencer. In addition, the sequencing procedure takes a day and more to generate the analysis results. The date should be represented by ISO 8601 as other date information.

8.7.2 Quality control metrics

The relevant quality control (QC) metrics for sequencing and analysis might be given. The report can include the overall QC metrics for biospecimen, the QC metrics for all variants, or the QC metrics for specific variants based on the report generator's decision.

Example: Sequencing yield, number of total reads, (average) read length (bp), number of reads mapped to reference genome (mapping yield, %), N (the bases were not used for base call) base (%), GC (%), Q20 (%), Q30 (%), On target coverage (%) > 1x, On target coverage (%) > 10x, n target coverage (%) > 20x, On target coverage (%) > 100x, mean depth, and uniformity

EXAMPLE GC = 53,2 %, AT = 47,8 %, Q30 = 94,4 %, Q20 = 97,1 %.

8.7.3 Base calling information

Information on base calling that is generated by base calling software for identifying a nucleotide sequence could be notified. The detailed fields will be as follows. The other fields can be added based on the report generator's decision.

8.7.3.1 Read depth

The average number of nucleotide contributing to a portion of an assembly could be reported as used in conventional bioinformatics fields.

EXAMPLE 100x.

8.7.3.2 Reference allelic depth

Allelic depth for the reference allele could be reported as used in conventional bioinformatics fields.

EXAMPLE 50x.

8.7.3.3 Alternative allelic depth

Allelic depth for the alternative allele could be reported.

EXAMPLE 50x.

8.7.3.4 Allele frequency

The frequency of alternative allele at each locus for each individual could be reported.

EXAMPLE 0,3, 30 %.

8.7.3.5 Genotype

The pair of allele presents at a single locus could be reported.

EXAMPLE AA, AC, AG, AT, CC, CG, CT, GG, GT, TT.

8.7.4 Sequencing platform information

Information on sequencing techniques and data including sequencing platform, capture method, and alignment algorithm should be given in text.

8.7.4.1 Type of sequencers

The specific sequencer that performs the sequencing should be given.

EXAMPLE Illumina Hiseq 2500, Thermo Fisher Ion Torrent, Illumina MiSeq.

8.7.4.2 Library preparation methods

The sequencing library preparation methods should be given.

EXAMPLE SureSelectXT.

8.7.4.3 Target capture methods

The exome or targeted region capture methods should be notified.

EXAMPLE Amplicon, probe capture.

8.7.4.4 Read type

Sequencing read type should be given.

EXAMPLE Single-end, Paired-end.

8.7.4.5 Read length

The sequencing read length information should be given.

EXAMPLE 101 bp, 35-250 bp.

8.7.5 Analysis platform information

The primary, secondary (if applicable tertiary) analysis pipelines should be mentioned. The parameter setting for pipelines should be also mentioned to confirm the reliability of variant call.

EXAMPLE GATK 3.5, CASAVA 1.7, Complete Genomics v2.2, Torrent Suite 5.0.2.

8.7.5.1 Alignment tools

The name of alignment tool and its version should be notified.

EXAMPLE BWA-MEM 0.7.12.

8.7.5.2 Variant calling tools

The name of variant calling tool and its version should be notified.

EXAMPLE GATK 3.5, SAMTools 1.3.1.

8.7.5.3 Other tools

EXAMPLE PICARD 1.9.3.

8.7.5.4 Chromosome coordinate system

The chromosome number can be started at 0 or 1. Therefore, there should be notified as the 0-based, 0-based half open, or 1-based coordinated system.

8.7.5.5 Annotation tools and databases

The name of annotation tools and source of databases that is publicly available or private should be reported.

EXAMPLE ANNOVAR (2016Feb01), SnpEff, 4.3, Ensembl v74.

8.8 References

All information should be cited by the proper references. References can be the published articles or curated database. Any consistent reference format can be used.

Annex A
(informative)

Example structure of clinical sequencing report

This annex demonstrates the composition of clinical sequencing report with the informative values.

Table A.1 — Summary

Fields[a]			Value[b]	Representation[c]
Clinical sequencing order	Clinical sequencing order code	Order code	14-RM-0000056	14-RM-0000056
	Date and time	Order date	2014–04–18	April 18, 2014
		Report date	2014–04–25	April 25, 2014
	Specimen information		13-S-048435_A1	13-S-048435_A1
Information on subject of care	Identifiers		12345678	12345678
	Name		Gildong Hong	Gildong Hong
Biomaterial information	Type of sample		CEN	Non-blood tissue
Genetic variations	Gene symbols and names		HGNC:1097, BRAF	BRAF
Recommended treatment	Medication		Vemurafenib, L01XE15 (ATC code)	Vemurafenib
Overall interpretation			Identified	Identified

NOTE The example of Summary part.

[a] The selected fields for Summary part.

[b] Code or value for each field.

[c] Representation in each field of the sequencing report.

Table A.2 — Detailed contents: Required

Fields[a]			Value[b]	Representation[c]
Clinical sequencing orders	Clinical sequencing order code	Order Code	14-RM-0000056	14-RM-0000056
		Information on sequencing order	OncoPanel_V2 (Targeted NGS for 505 genes, T/N pair)	OncoPanel_V2 (Targeted NGS for 505 genes, T/N pair)
	Date and time	Order date	2014–04–14	April 14, 2014
		Specimen collection date	2014–04–17	April 17, 2014
		Order received date	2014–04–25	April 25, 2014
		Report date	2014–05–08	May 08, 2014
	Specimen information		13-S-048435_A1	13-S-048435_A1
Information on subject of care	Identifiers		12345678	12345678
	Name		Gildong Hong	Gildong Hong
	Birth Date		1947–04–29	April 29, 1947
	Sex		1	Male
	Ethnicity		2040–04–01	Korean
Information of legally authorized person ordering clinical sequencing			D060001	Chulsoo Kim
Performing laboratory	Basic information		AMC Genomic Pathology Lab	AMC Genomic Pathology Lab, 02–3010–8460
	Information of report generator		D110001	Min Lee
	Information of legally confirmed person on sequencing report		D030001	Sejin Kim
Associated diseases and phenotypes			C34.90	C34.90 Malignant neoplasm of unspecified part of unspecified bronchus or lung
Biomaterial information	Type of sample		CEN	Non-blood tissue
	Genomic source class in biomaterial		2	Somatic
	Conditions of specimen		Acceptable, Tumor proportion of the tested tissues (%): 90	Acceptable, Tumor proportion of the tested tissues (%): 90

NOTE The example of detailed part (required fields).

[a] The fields for required fields.

[b] Code or value for each field.

[c] Representation in each field of the sequencing report.

Table A.2 *(continued)*

Genetic variations	Gene symbols and names		HGNC:1097, BRAF	BRAF
	Sequence variation information	HGVS nomenclature	c.1799T > A_p.V600E	c.1799T > A_p.V600E Kinase domain (exon 15)
		Effects of variants	Substitution (missense)	Substitution (missense)
		Sequence variant ID	COSM476	http://cancer.sanger.ac.uk/cosmic/mutation/overview?id=476
Classification of variants	Pathogeny		Pathogenic	Tier 1 (Pathogenic, Identified)
	Clinical relevance		Identified	
Recommended treatment	Medication		Vemurafenib, L01XE15 (ATC code)	Vemurafenib
	Clinical trial information		NCT01791309	Vemurafenib and Panitumumab Combination Therapy in Patients with BRAF V600E Mutated Metastatic Colourectal Cancer (https://clinicaltrials.gov/ct2/show/NCT01791309)

NOTE The example of detailed part (required fields).

[a] The fields for required fields.

[b] Code or value for each field.

[c] Representation in each field of the sequencing report.

Table A.3 — Detailed contents: Optional

Fields[a]		Value[b]	Representation[c]
Medication history		N/A	N/A
Family history/Pedigree information		N/A	N/A
Reference genome version		GRCh37.p13	GRCh37 Patch Release 13 (Released June 28, 2013)
Racial genomic information		1000 Genomes	1000 Genomes
Genetic variations	Gene symbols and names	HGNC:11998, TP53 HGNC:17278, PNRC1 HGNC:30988, BNC2	TP53 PRNC1 BNC2
	HGVS version	HGVS Version 15.11	HGVS Version 15.11

NOTE The example of detailed part (optional fields).

[a] The selected fields for optional fields.

[b] Code or value for each field.

[c] Representation in each field of the sequencing report.

Table A.3 *(continued)*

Detailed sequencing information	Clinical sequencing Date		2014–05–04	May 04, 2014
	QC metrics	Number of total reads	6950582	6,950,582 bp
		Mean target coverage	121.2	121.2 X
		% targets bases covered 100X	93.1	93.1 %
	Sequencing platform	Type of sequencers	MiSeq	Illumina MiSeq
		Library preparation methods	SureSelectXT Reagent Kit, HSQ 96	SureSelectXT Reagent Kit, HSQ 96
		Target capture methods	SureSelect Biotinylated RNA library "Baits" (SureSelectXT Custom panel)	SureSelect Biotinylated RNA library "Baits" (SureSelectXT Custom panel)
		Read type	pair-ended	Paired Ends
		Read length	150	150 bp (75x2)
	Analysis platform information	Alignment tools	BWA	BWA-MEM 0.7.12
		Variant calling tools	SNV (Mutect2) indel (somatic indellocator) CNV (ReCapSeg) Structural variation (BreakMer 0.0.2)	SNV (Mutect2) indel (somatic indellocator) CNV (ReCapSeg) Structural variation (BreakMer 0.0.2)
		Chromosome coordinate system	one-based	one-based
		Annotation tools and databases	VEP: Variant Effect Predictor version 85	VEP: Variant Effect Predictor version 85

NOTE The example of detailed part (optional fields).

[a] The selected fields for optional fields.

[b] Code or value for each field.

[c] Representation in each field of the sequencing report.