Questionnaire to survey availability and define the gaps in the field of metadata standards and procedure(s) for genome-sequencing processes and data commissioned by the data management work group of the Personalised Medicine (PM) programme of ZonMw, Zilveren Kruis and KWF.

**Contact details**

Name:
Specialism/function:
e-mail:
Tel:
Re: ZonMw PM project(s):

**Objectives and activities involved in surveying metadata standards**

The ZonMw data management work group aims to improve the quality of data management in the framework of Personalised Medicine in Dutch healthcare and more specifically to optimise the outcomes of genome-sequencing research. To improve the quality of data management in genome-sequencing research, the work group has set two objectives:

1. Developing an accepted, widely supported standard /procedure for data management in genome-sequencing research.
2. Draw the attention of the field to this standard so all researchers are aware of its existence and its relevance and apply it where appropriate.

In the coming period we shall be working on firming up the details of frameworks and criteria for a subsidy round aimed at the development of the specified data/metadata standards and procedures for sequencing processes and data in the context of research into care applications in rare diseases and oncology. An initial step in that direction involves making an inventory of the availability of metadata standards (public and private) for genome-sequencing processes and data (this also concerns data for HTA: Health Technology Assessment).

As one of the experts in this specialism, we would like to interview you and possibly your colleagues. Prior to the interview we would like to obtain an idea of what data and metadata are generated in your organization and how they are recorded. This will help us to structure the interview. In preparation for the interview, we have therefore created the questionnaire below. The questions specifically concern: workflow, type of data/metadata, data/metadata generation, management, quality, exchange and integration. NB: text in italics contains clarification or examples. It is up to you to fill in the relevant information for you, or include it as an attachment.

Could you or someone else please complete this questionnaire as thoroughly as possible and return it to us by the latest two days before the interview? We would like to ask you especially to indicate where you see gaps in the availability of standards/procedures.

If you have any questions in the meantime, feel free to contact:

- Jeroen Beliën, jam.belien@vumc.nl, mobile phone +31 6 10 736 072
- Anke Kip, anke.kip@lygature.org, mobile phone +31 6 18 04 39 13

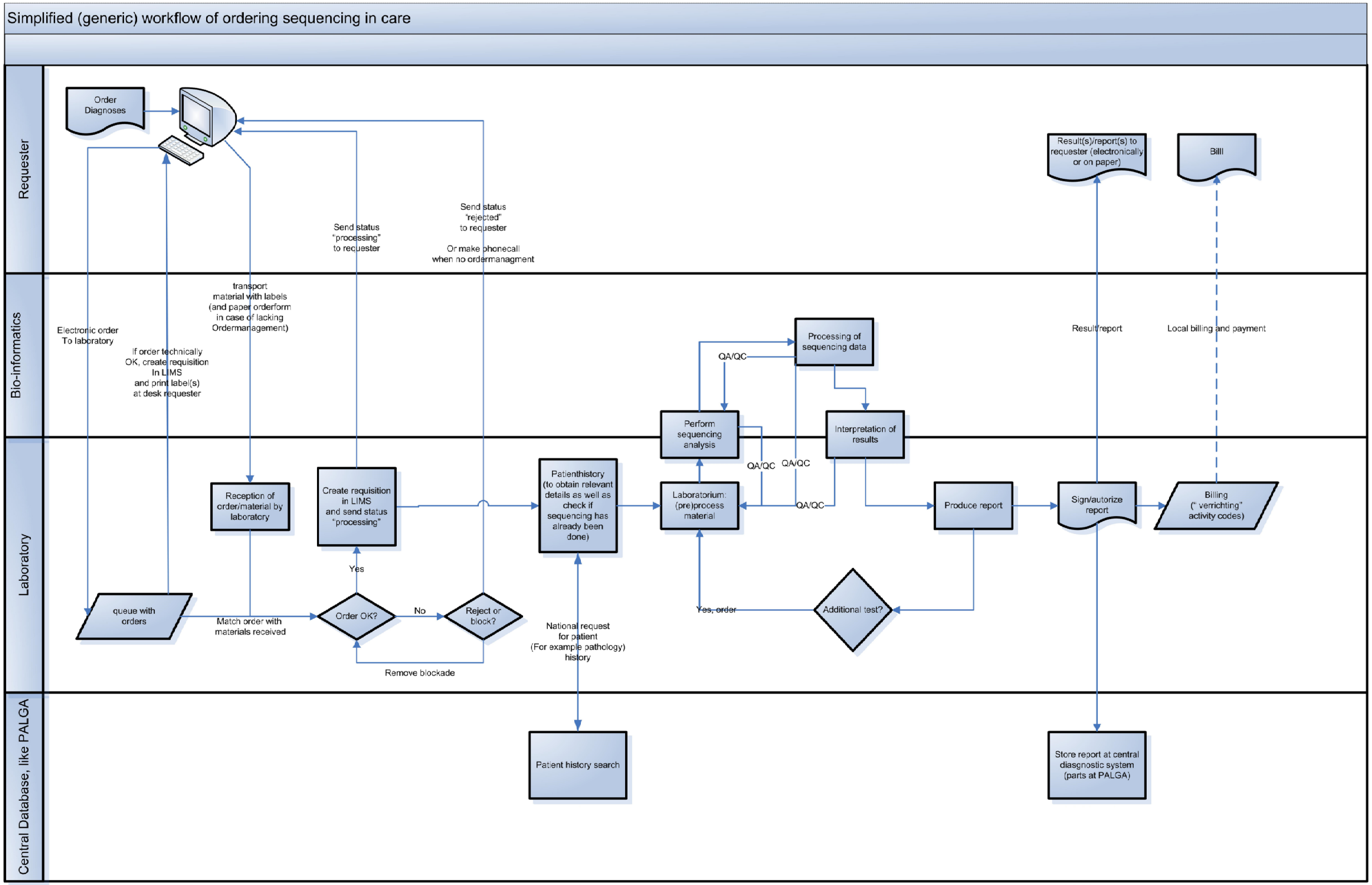
Thank you in advance for your cooperation.

Jeroen Beliën and Anke Kip, on behalf of the ZonMw data management work group:

Dr. Morris Swertz, chair,
Dr. Jan-Willem Boiten,
Dr. Rob Hooft,
Dr. Margreet Bloemers,
Dr. Ana de Castro,
Dr. Salome Scholtens,
Ilse Custers.

| | | |
|---|---|---|
| **Respondent(s)** | | |
| 1 | Which person(s) within your organisation, or affiliated organisation(s), completed this questionnaire? | *If only you completed it, then the contact details given above are sufficient. If other people were involved/consulted, please add their details here and don't forget to specify their "specialism/function" (eg data manager, data scientist, doctor, lab technician, health technology assessment specialist, bio-computer scientist).*<br><br>*Name:*<br>*Specialism/function:*<br>*e-mail:*<br>*Tel:*<br>*Regarding project(s):* |
| 2 | Are there other people and/or companies and/or authorities outside your organisation that you consider experts who could help us with the inventory by completing this questionnaire (and possibly an interview)? | *If you know of others who could or should complete this questionnaire, please give their contact details below:*<br><br>*Name:*<br>*Specialism/function:*<br>*e-mail:*<br>*Tel:*<br>*Regarding project(s)/expertise:*<br><br>*Name:*<br>*Specialism/function:*<br>*e-mail:*<br>*Tel:*<br>*Regarding project(s)/expertise:*<br><br>*[...]* |
| **Image board: process overview (see next page)** | | |
| 3 | a) Is that image board complete concerning you and your activities involving data and workflow (logistics, processes)?<br>b) Can you indicate on this image board where your activities are located?<br>c) Can you indicate whether the image board below differs or should differ for research compared to care?<br><br>If you would like an explanation of this image board, contact us directly. | *If the workflow you use or would like to follow does not match, please state how and where it deviates from the one below (original Microsoft Visio file is enclosed, but you can also print out the image board and make your changes on the paper).*<br><br>*If you restrict yourself to a certain part of the process and would like it described in more detail to support your answers better, we ask you to sketch or submit that particular part of the process in a similar way in the detail you desire.* |
| 4 | Within this displayed image board, it is possible that at one or more moments a specific consent (e.g. informed consent) is requested, recorded, accessed, checked, or given as part of a process step. Could you state which applies according to you (should apply), who is involved or required, and at what times? Specify differences between care and research. | *You do not have to cover the entire legal framework, just describe which steps, documents, etc. are essential for being able to conduct the entire process and which parts of it should be accessible for you, and which you record/change yourself (distinguish between genome sequencing as part of care or for a research project).* |

This questionnaire is based on the envisaged high-level, simplified, generic data and workflow flowchart below for ordering sequencing in care: “the image board” (a high resolution version is supplied as a separate file)


Simplified (generic) workflow of ordering sequencing in care
Requester
Bio-informatics
Laboratory
Central Database, like PALGA
Order Diagnoses
Electronic order To laboratory
If order technically OK, create requisition In LIMS and print label(s) at desk requester
transport material with labels (and paper orderform in case of lacking Ordermanagement)
Send status “processing” to requester
Send status “rejected” to requester
Or make phonecall when no ordermanagment
queue with orders
Match order with materials received
Reception of order/material by laboratory
Create requisition in LIMS and send status “processing”
Yes
Order OK?
No
Reject or block?
Remove blockade
Patienthistory (to obtain relevant details as well as check if sequencing has already been done)
National request for patient (For example pathology) history
Patient history search
Laboratorium: (pre)process material
Perform sequencing analysis
QA/QC
Processing of sequencing data
QA/QC
QA/QC
Interpretation of results
QA/QC
Additional test?
Yes, order
Produce report
Sign/autorize report
Result/report
Result(s)/report(s) to requester (electronically or on paper)
Store report at central diasgnostic system (parts at PALGA)
Billing (“ verrichting” activity codes)
Local billing and payment
Billl

LIMS: Laboratory Information Management System
PALGA: Pathologisch Anatomisch Landelijk geautomatiseerd archief
QA/QC: check result, if necessary perform earlier step(s), etc.

| | | |
|---|---|---|
| **Type of data & data generation** | | |
| 5 | Which data/metadata are collected as part of your work in each process workflow step? | ***NB:*** *by collection of data, we are referring to the primary data like:*<br>- *Clinical data*<br>- *Data from samples and derivatives*<br>- *Raw or processed data of genome sequencing*<br>- *Cost data (e.g. cost of consumables, personnel cost)*<br><br>*The collection of data also refers to additional data such as: who records them, where are they recorded, what device or what tool (which version) is used, when is this done (date and time), and how are the genome data generated (for example, if a protocol is followed for reprocessing material, then specify the protocol; if raw data are processed and analyzed with software/pipelines, then specify which one(s))*<br><br>*Thus, for each process/workflow step, please specify:*<br>• *What (which data)*<br>○ *Which data must be recorded (are mandatory)? (without these data neither you nor the next person in the chain can proceed)*<br>○ *Which data are important for you, but you are not obliged to record them?*<br><br>*Note: By derivative we mean a product derived from a sample: for example, DNA from blood, DNA from tissue.* |
| 6 | How are these data/metadata recorded? | • *Who records the data/metadata?*<br>• *Where are the data recorded? (eg wherever the data are generated (at the source), or later)*<br>• *Which system is used to record the data/metadata?*<br>• *Are the data/metadata automatically saved?* |
| **Data/sample quality and management** | | |
| 7 | For the data from question 5 (if possible for each step in the previously given workflow), specify:<br>1) In what format or according to what standard are data/metadata recorded<br>2) Which standard nomenclature /ontology/thesaurus/ codebook is used | *Clarification: Along with the requested details about data in question 5, it is relevant to inventory, for example, how the quality of your data are guaranteed, how others are capable of understanding and interpreting your data, etc. With the questions given on the left, we try to survey that.*<br><br>*7.1 by format we mean, for example, a flat text file, comma separated values, VCF, BAM, BED, BGEN, HFD5, MINISEQE, etc.*<br>*7.2. are the data/metadata recorded according to a certain standard or ontology (eg Dublin Core, SNOMED-CT, ICD10, LOINC, …)* |
| 8 | For the data from question 5 (if possible for each step in the previously given workflow) specify:<br>1) If relevant, who checks the given data for correctness, completeness, etc.<br>a. If you receive material or data<br>b. If you record/change/delete data<br>c. …<br>2) Who manages the data | *8.1 Who: this can be you, or a colleague, a group, or an agency. An example of a check could be: Date of receipt of material cannot be later than the date of authorization of a final report*<br>*8.2 eg you can manage the data on the PC on your desk, or the ICT department can do that, or a service provider can do that.* |
| 9 | Are data, materials collected according to SOPs? | *If several centres/departments are involved in the analysis: is the same SOP used?* |
| 10 | Is quality control or validation used for:<br>• Bodily materials<br>• Use of control samples/material (calibration) | *If yes, give description*<br>*eg validation that*<br>• *Tissue block/slide at least 75% tumour cells*<br>• *[…]* |

| | | |
|---|---|---|
| 11 | To promote quality control, are so-called rounds being done in which everyone has to analyze the same sample? | *If yes, which type and to what purpose? Where are the results stored, or how are they reported?* |
| **Other info data items/sets** | | |
| 12 | For each stated data item/set, indicate whether the data is isolated or coupled/linked. | *eg you use a participant-ID to identify a participant, usually this participant-ID is coupled by someone else to a person/patient-ID: the data is then coupled; if you have access only to an anonymous participant-ID or experiment-ID, then the data is isolated.*<br>*Another example: if you use gene-identifiers like those managed by http://www.ensembl.org/, your data (in this case the gene) is then linked.*<br><br>*If someone else has the key, who are they and following what protocol/procedure can access to the key be granted?* |
| 13 | For each stated data item/set, indicate whether these data are static or dynamic. | *Static data never or very rarely change; dynamic data are often subject to change.*<br>*It is important to know whether data are static or dynamic in connection with, for example, version management or whether certain static data serve as the source or can be discarded because they are fully reproduceable and can be produced again.* |
| 14 | For each stated data item/set, indicate whether you:<br>- Store the data/make it available after the end of the project and to whom?<br>- Intend to/must destroy the data? | *Please specify for both questions why you do/do not do this.* |
| 15 | What assumptions are involved and what are the consequences for recording data/metadata? | *For example:*<br>- *only human, only mouse, …*<br>- *only germline, only somatic, a combination*<br>- *which genome sequencing technique(s) you apply: WGS, exome sequencing, ….*<br>*State if necessary:*<br>- *any assumptions for each data(item).*<br>- *What differences you can identify given that you are working with one procedure and certain assumptions which may differ from those of other researchers/technicians (eg are there essential differences in recording data/metadata in the analysis of germline or somatic cells, and what are they)* |
| 16 | What annotations of the data are saved and where and whom are they intended for? | *eg interpretation of the data (possibly in the form of a report to applicant), scores, illustrations of genome sequencing result* |

| | | |
|---|---|---|
| **Data exchange and integration** | | |
| 17 | Are the data/samples generated only by your own department/institution or in a partnership?<br>If part of the process takes place in another institution or in another department in your organization, which data are required to allow the other to carry out their part of the process and report back on the data/results? | *We want to survey here what data/information is essential within the process for correct conduct and thus must always be recorded completely and correctly ("the real minimum data of and within the entire process chain")* |
| 18 | Is the tool/software used for storing/recording the data/metadata capable of:<br>• Exchanging data with other tools and/or software packages?<br>• Importing already existing genome sequencing data (possibly from others)?<br>• Exporting data to other packages (possibly genome sequencing) or formats | |
| 19 | Do you use software that can bind or link different data together? | *If yes, which one(s)?* |
| **Identifying known gaps** | | |
| 20 | Can you specify where you see gaps (please also clarify why):<br>- In recording, searching, managing, publishing, etc. data/metadata for/about genome sequencing?<br>- In setting up or working with a procedure for genome sequencing? | |
| **Which other initiatives / projects are you involved in where genome sequencing plays a role** | | |
| 21 | *Can you state whether you are involved in other genome-sequencing (implementation) projects/assignments/etc.?* | *If yes, do these projects/assignments/etc. use the same processes and/or data/metadata standards and which ones* |
| **Additional requirements / comments, etc.** | | |
| 22 | *Do you have any further comments/points to make that were not covered in this questionnaire, and if so, what are they?* | |