The Ten Commandments of Translational Research Informatics

Tim HULSEN[a,1]

[a]Department of Professional Health Solutions & Services, Philips Research, Eindhoven, The Netherlands

[1]Corresponding author. E-mail: tim.hulsen@philips.com. ORCID: 0000-0002-0208-8443

Abstract. Translational research applies findings from basic science to enhance human health and well-being. In translational research projects, academia and industry work together to improve healthcare, often through public-private partnerships. This "translation" is often not easy, because it means that the so-called "valley of death" will need to be crossed: many interesting findings from fundamental research do not result in new treatments, diagnostics and prevention. To cross the valley of death, fundamental researchers need to collaborate with clinical researchers and with industry so that promising results can be implemented in a product. The success of translational research projects often does not only on the fundamental science and the applied science, but also on the informatics needed to connect everything: the translational research informatics. This informatics should enable the researchers to store their 'big data' in a meaningful way, to ensure that results can be analyzed correctly and enable application in the clinic. This translational research informatics field has overlap with areas such as data management, data stewardship and data governance. The author has worked on the IT infrastructure for several translational research projects in oncology for the past nine years, and presents his lessons learned in this paper in the form of ten commandments. These commandments are not only useful for the data managers, but for all involved in a translational research project. Some of the commandments deal with topics that are currently in the spotlight, such as machine readability, the FAIR Guiding Principles and the GDPR regulations, but others are not mentioned often in publications around data stewardship and data management, although they are just as crucial for the success of a translational research project.

Keywords: translational research, medical informatics, data management, data curation, data science

1. Introduction

Translational research applies findings from basic science to enhance human health and well-being. In a medical research context, it aims to "translate" findings in fundamental research into medical practice and meaningful health outcomes. In translational research projects, academia and industry work together to improve healthcare, often through public-private partnerships [1]. This “translation” is often not easy, because it means that the so-called “valley of death” [2] will need to be crossed: many interesting findings from fundamental research do not result in new treatments, diagnostics and prevention. To cross the valley of death, fundamental researchers need to collaborate with clinical researchers and with industry so that promising results can be implemented in a product. Examples of initiatives supporting translational research are EATRIS [3], the European Infrastructure for Translational Medicine and NCATS [4], the National Center for Advancing Translational Sciences in the USA.

The success of translational research projects often does not only on the fundamental science and the applied science, but also on the informatics needed to connect everything: the ‘translational research informatics’. This type of informatics was first described in 2005 by Payne et al. [5], as the intersection between biomedical informatics and translational research. Translational research informatics should enable the researchers to store their ‘big data’ in a meaningful way, to ensure that results can be analyzed correctly and enable application in the clinic [6]. This translational research informatics field has overlap with areas such as data management, data stewardship and data governance, which are increasingly getting attention recently. Data management (in research) is the care and maintenance of the data that is produced during the course of a research cycle. It is an integral part of the research process and helps to ensure that your data is properly organized, described, preserved, and shared [7]. Data management ensures that the story of a researcher’s data collection process is organized, understandable, and transparent [8]. According to Rosenbaum, 2010 [9], data stewardship is a concept with deep roots in the science and practice of data collection, sharing, and analysis. It denotes an approach to the management of data, particularly data that can identify individuals. Data governance is the process by which responsibilities of stewardship are conceptualized and carried out. Perrier et al. (2017) presents an extensive overview of 301 articles on data management, distributed over the six phases of the Research Data Lifecycle [10]: (1) Creating Data, (2) Processing Data, (3) Analysing Data, (4) Preserving Data, (5) Giving Access to Data and (6) Re-Using Data. It shows that most publications focus on phases 4 to 6 (especially phase 5), but not so much on phases 1 to 3, while these are equally important.

The author has worked on the IT infrastructure, data integration and data management for several translational research projects in oncology [11-15] for the past nine years, as well as the Dutch translational research informatics project CTMM-TraIT [16], and presents his lessons learned in this paper in the form of ten commandments. These commandments are not only useful for the data managers, but for all involved in a translational research project, since they touch upon very crucial elements such as data quality, data access and sustainability, and cover all phases of the research data lifecycle. As opposed to most publications in the field of data management, this article even puts an emphasis on the early phases of the research data lifecycle.

2. The Ten Commandments

Commandment 1: Create a separate Data Management work package

When clinicians, biologists and other researchers come together in a translational research project, they often do not think about data management, data curation, data integration and the IT infrastructure, except for when it is already too late: the data sit on several computers scattered over different organizations, and nobody knows how to combine them and make sense of them. The solution: create a

separate work package (WP) or work stream (WS) on data management. This WP can be thought of as a sub-project, which has its own goal (or even a list of milestones and deliverables) and has FTEs and other financial resources allocated. Within this WP, a data management plan (DMP) will be created which describes exactly how the data management will take place (such a DMP is obligated nowadays in several funding programmes such as Horizon 2020 [17], and for good reason). Since this data management WP will have touchpoints with the other (data-generating) work packages, the data management WP leader needs to be involved in all project meetings. It is also advised to create a separate WP on data analysis, which gets its input from the data management / data integration WP (Figure 1).

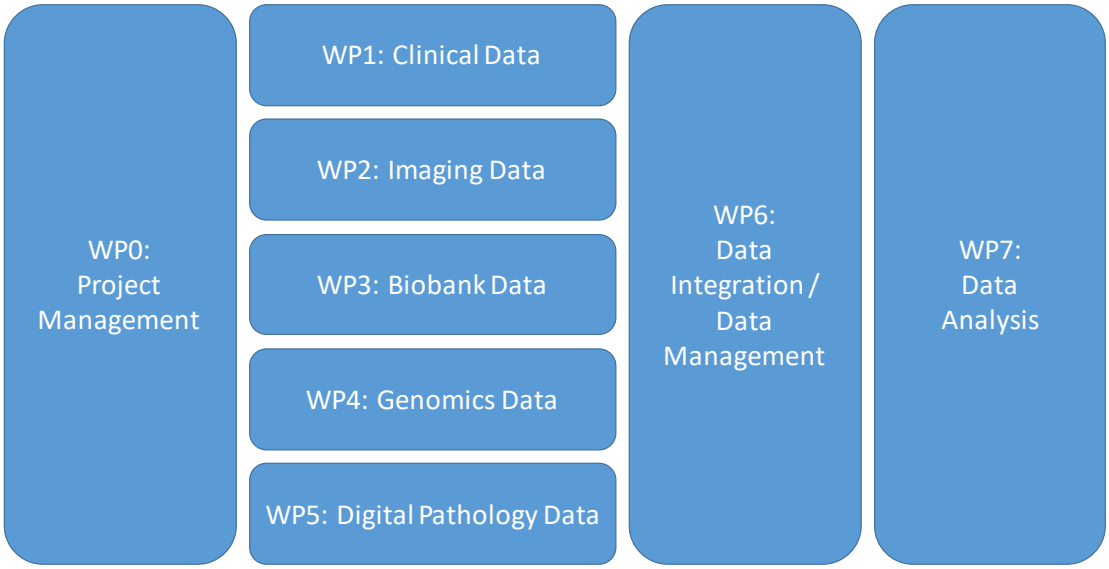

Figure 1. A proposed WP structure for a translational research project

Commandment 2: Reserve time and money for data entry

The Principal Investigators of a translational research project like to talk about the grand scheme of things: scientific hypotheses, great breakthroughs and publications in big journals, but often forget that they need people to do the 'dirty job'; for example, the entering of clinical data into an Electronic Data Capture (EDC) system such as OpenClinica [18] or Castor EDC [19] or REDCap [20], as part of WP1 in figure 1. This work often is assigned to trial nurses, who usually already have enough, more pressing, things on their hands. The data entry work is on the bottom of their priority list, which can cause delays and even errors. Which is really a concern, because data quality is a very important matter when it comes to data analysis [21]. Even when the data is automatically extracted from an EMR and entered into the EDC system, this needs to be checked by someone. The term “Garbage in, garbage out” (GIGO) comes from computer science, but applies to medical data as well [22]. The solution here is to reserve money to hire people who can do this job for a certain amount of hours per week. By spending relatively little money on data entry, one can save a lot of time and money by not having to redo analyses because of missing or erroneous data.

This commandment does not apply to just clinical data but to the other data types as well: imaging and digital pathology data often need to annotated, biobank data usually needs to be exported from a Laboratory Information Management System (LIMS), and raw genomics data needs to be processed first before it can be used in an integrated database. All these steps need sufficient resources to make sure that they are carried out correctly.

Commandment 3: Define all data fields up front together with the help of data analysis experts

Within the Prostate Cancer Molecular Medicine (PCMM [13]) project, we noticed after a few years that some information that was essential to answer certain research questions was not being collected in the electronic case report form (eCRF). A second eCRF needed to be constructed, which resulted in a lot of time being spent on going back to the patient's entries in the hospital system and collect the data, if it was there at all. We learned our lesson. Within the Movember GAP3 project [14], all parties together created the Movember GAP3 codebook, which was an extensive list of data fields designed to answer all research questions that we could think of at the start of the project. The statisticians within the project were very much involved in the codebook creation, because they had the clearest insights on what was needed here. Besides the data field name, we stored the data type (integer, float, string, categorical, etc.) and the unit (days, years, cm, kg, kg/m2, ng/ml, etc.), and (in case of a categorical data type) listed the categories (e.g. the TNM staging system). In case of a derived data field, the codebook explains how this data field is calculated. Examples here are data fields such as age, BMI and days since diagnosis. Because part of the Movember GAP3 data was retrospective, we created some data model mapping scripts [23] to map the existing data to this codebook, as well as some data curation scripts that check if the data falls into the expected range and does not contain any discrepancies.

Figure 2 describes the clinical data collection process. The green parts are these steps of the process that are necessary, whereas the red parts are steps that are unnecessary in modern translational research. The green parts include the construction of the codebook, the eCRF creation, the data entry into the eCRF, the data integration into the database and the data analysis leading to results. The red parts include the creation of and the data entry into the paper CRF, which ideally should be avoided because this gives the data entry person double work. The information should be entered directly into the eCRF instead. The yellow line should only be followed when, even after carefully constructing the codebook, more data fields need to be included. Ideally, the codebook should also be compliant with ontologies for translational research such as the Basic Formal Ontology (BFO), the ontologies listed by the Open Biological and Biomedical Ontology (OBO) foundry and the Relation Ontology (RO) [24].

Other data types that are important in translational research, as listed in figure 1 (WP 2-5), often do not need a codebook because they are stored in standardized file formats such as DICOM (for imaging and digital pathology data). However, information derived from these data types, such as PI-RADS scores (from prostate cancer MR images) and Gleason scores (from prostate cancer digital pathology images) should be stored in the codebook as well, to ensure that these values can be compared with the clinical data gathered in the eCRF.

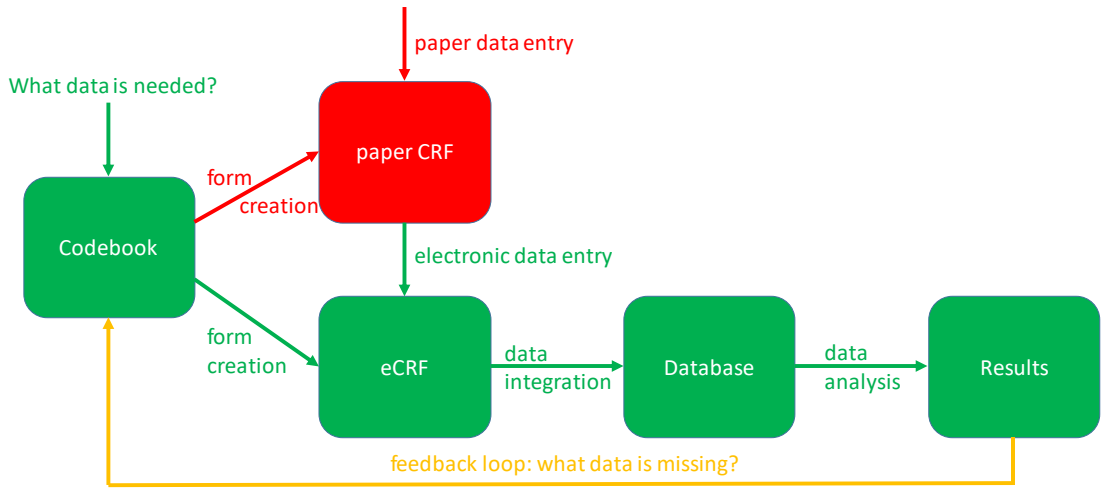

Figure 2. The clinical data collection process.

Commandment 4: Make clear arrangements about data access

In large consortia, especially consortia with both academic partners and commercial partners, data access can be a sensitive issue. That's why it needs a clear arrangement up front. Of course, data access needs to be arranged in the informed consent as well, as patients are the data owners, and the General Data Protection Regulation (GDPR [25], within the EU) and the Health Insurance Portability and Accountability Act (HIPAA [26], within the US) have strict regulations about the patient's privacy. The GDPR puts some constraints on data sharing, e.g., if a data controller wants to share data with a third party, and that third party is a processor, then a Data Processor Agreement (DPA) needs to be made. Also, the informed consent that the patient signs before participating in a study, needs to state clearly for what purposes their data will be used.

Ideally, at the end of the project, when all goals have been met and results have been published, the de-identified data should be shared with the whole world, if privacy regulations allow it. After all, the goal of a translational research project is to "translate" findings in fundamental research into medical practice and meaningful health outcomes, which can only be achieved if data is being shared as soon as possible, because then the whole world can use the data. This future public availability of the data should be included in the informed consent as well. As GDPR article 4 [25] states that “consent of the data subject means any freely given, specific, informed and unambiguous indication of the data subject’s wishes by which he or she, by a statement or by a clear affirmative action, signifies agreement to the processing of personal data relating to him or her”, this means that there should be a “yes/no” question or checkbox for the option to share the data in a public repository at the end of the study. If the patient answers “no” or does not check the box, the patient should still be allowed to enter the study, but his/her data cannot be submitted to a public repository.

Commandment 5: Agree about de-identification and anonymization

The responsibility for proper de-identification of the data often lies with the organization that collects the data (usually the hospital), because they are the ones that have the Electronic Health Record (EHR). They should create a study subject ID for each subject, which can only be mapped to the original subject ID by a mapping table residing at the hospital. If the party performing the data integration receives data that is not properly de-identified by the hospital, it should be destroyed immediately, and a new data submission should be requested. If a subject at any time requests to have his/her data removed from the central database, the hospital should inform the data manager about which data belonging to which study subject ID needs to be removed. In the case that the data integration expert also needs to do the de-identification and anonymization, this should be arranged very clearly in the informed consent and the data processor agreement. For textual and numerical data, open-source software packages are available that can help with anonymization, such as the ARX anonymization tool [27]. For imaging data, anonymization tools are available that can strip any identifiable information from Digital Imaging and Communications in Medicine (DICOM) tags, such as the DICOM anonymizer [28] and the DicomCleaner [29].

Commandment 6: Reuse existing software where possible

There is usually no need to develop tools for data capturing, data management, data quality control, etc., because there are open source tools available for this. Within the Translational Research IT (TraIT) project [16] of the Center for Translational Molecular Medicine (CTMM), a list of suitable open source tools was created, which included OpenClinica [18], XNAT [30] and tranSMART [31]. For areas where there was no tool available, software was created. An overview of the TraIT tools can be found at https://trait.health-ri.nl/trait-tools/. Most translational research projects have similar problems, so when

starting a new project, it is generally a good idea to see how they solved these problems, and if their solution can be reused. This reusability also increases the reproducibility of the research, because there is no reliance on obscure, custom-made computer scripts or websites. Table 1 shows an up-to-date list of freely available software related to translational research, including a description and the main data type it processes. This list was created by combining the above mentioned overview with an extensive PubMed search.

Name	**Main data type**	**Description**	**URL**
cBioPortal [32]	Genomics	The open source cBioPortal for Cancer Genomics provides visualization, analysis, and download of large-scale cancer genomics data sets. A public instance of cBioPortal (https://www.cbioportal.org) is hosted and maintained by Memorial Sloan Kettering Cancer Center. It provides access to data by The Cancer Genome Atlas as well as many carefully curated published data sets. The cBioPortal software can be used to for local instances that provide access to private data.	https://github.com/cBioPortal/
Dicoogle [33]	Imaging	Dicoogle is an open source Picture Archiving and Communications System (PACS) archive. Its modular architecture allows the quick development of new functionalities, due the availability of a Software Development Kit (SDK).	http://www.dicoogle.com/
Galaxy [34]	Genomics	Galaxy is a scientific workflow, data integration, and digital preservation platform that aims to make computational biology accessible to research scientists that do not have computer programming or systems administration experience. Although it was initially developed for genomics research, it is largely domain agnostic and is now used as a general bioinformatics workflow management system	https://usegalaxy.org/
I2B2 [35]	Clinical	Informatics for Integrating Biology and the Bedside (i2b2) is one of the sponsored initiatives of the NIH Roadmap National Centers for Biomedical Computing. One of the goals of i2b2 is to provide clinical investigators with the software tools necessary to collect and manage project-related clinical research data in the genomics age as a cohesive entity; a	https://www.i2b2.org/

		software suite to construct and manage the modern clinical research chart.	
Occhiolino [36]	Biobank	GNU LIMS, also known as Occhiolino is an open source laboratory information management system (LIMS), aiming for healthcare laboratories as its fully compatible with GNU-Health with complete workflow process control integration.	http://lims.gnu.org/
OpenClinica Community Edition [18]	Clinical	The world's first commercial open source clinical trial software serving for the purpose of clinical data management (CDM) and electronic data capture (EDC).	https://www.openclinica.com/
OpenSpecimen [37]	Biobank	The OpenSpecimen LIMS application allows bio-repositories to track biospecimens from collection to utilization across multiple projects, collect annotations, storage containers, track requests and distribution, and has multiple reporting options. It streamlines management across collection, consent, QC, request and distribution and is highly configurable and customizable.	https://www.openspecimen.org/
Orthanc [38]	Imaging	Orthanc aims at providing a simple, yet powerful standalone DICOM server. It is designed to improve the DICOM flows in hospitals and to support research about the automated analysis of medical images.	https://www.orthanc-server.com/
QuPath [39]	Digital pathology	QuPath is new bioimage analysis software designed to meet the growing need for a user-friendly, extensible, open-source solution for digital pathology and whole slide image analysis.	https://qupath.github.io/
REDCap [20]	Clinical	REDCap (Research Electronic Data Capture) is a browser-based, metadata-driven EDC software solution and workflow methodology for designing clinical and translational research databases. Development of the software takes place by collaborative software development through the REDCap consortium.	https://projectredcap.org/
SlideAtlas [40]	Digital pathology	SlideAtlas is an open-source, web-based, whole slide imaging platform. It provides features for multiple stages of a digital pathology workflow, including automated image uploading, image organization and management, automatic alignment and	https://slide-atlas.org/

		high-performance viewing of 3D image stacks, online annotation/markup and collaborative viewing of images.	
tranSMART [31]	Integration	tranSMART is a suite of data exploration, visualization, and ETL tools, which were originally developed by Johnson & Johnson for translational research studies. The software was released in 2012 as an open-source platform. It continues to be developed and maintained by a community effort, coordinated by the i2b2 tranSMART Foundation.	https://transmartfoundation.org/current-transmart-platform-release/
XNAT [30]	Imaging	XNAT is an open-source imaging informatics software platform dedicated to helping perform imaging-based research. XNAT's core functions manage importing, archiving, processing and securely distributing imaging and related study data.	https://www.xnat.org/

Table 1. Freely available software in the area of Translational Research Informatics.

Commandment 7: Make newly created software reusable

Although it is proposed at commandment 6 that existing software should be reused as much as possible, there might be cases where study-specific software needs to be created, for example to perform novel analyses. If there are no intellectual property issues, this newly created software can be submitted to repositories such as GitHub [41], SourceForge [42] or FigShare [43], or it can be made available on a custom-made website, and then referenced on Zenodo [44]. Griffin et al. [45] gives a good overview of the possibilities. This way, future translational researchers can reuse your software and don't need to reinvent the wheel. Github already hosts several scientific data management packages, such as Rucio (https://github.com/rucio/rucio), ISA tools (https://github.com/ISA-tools/isa-api) and Clowder (https://github.com/ncsa/clowder). There is also an overview of all 1,720 bioinformatics repositories on GitHub available [46]. If the software is submitted to one of these popular repositories, and it is accompanied with metadata that describes accurately what the software can do, it will be much easier for researchers to find the software and share it with other potential users.

Commandment 8: Adhere to the FAIR Guiding Principles

In 2016, the FAIR Guiding Principles for scientific data management and stewardship [47] were published. FAIR stands for the four foundational principles - Findability, Accessibility, Interoperability, and Reusability - that serve to guide data producers and publishers as they navigate around the obstacles around data management and stewardship. The difference with similar initiatives is that the FAIR principle do not only support the reuse of data by individuals, but also put emphasis on enhancing the ability of machines to automatically find and use the data. The elements of the FAIR Guiding Principles are related, but independent and separable:

- Findability is about making sure that the data can be found, e.g. by using a unique and persistent identifier and by the use of rich metadata which is registered or indexed in a searchable resource.

- Accessibility refers to the retrievability of the data and metadata by their identifier using a standardized communications protocol, and the access to the metadata even when the data are no longer available.
- Interoperability is about the usage of ontologies, vocabularies and qualified references to other (meta)data so that the data can be integrated with other data.
- Reusability refers to describing the (meta)data with a plurality of accurate and relevant attributes, releasing with a clear and accessible data usage license, etc., in order to enable reuse of the data.

The FAIR Guiding Principles should be applied to both data and software created in a translational research project, to achieve transparency and scientific reproducibility. An example of a FAIR-compliant dataset, is the Rembrandt brain cancer dataset [48]. This dataset is 'findable': it is hosted in the Georgetown Database of Cancer (G-DOC), with provenance and raw data available in the National Institute of Health (NIH) Gene Expression Omnibus (GEO) data repository. These resources are publicly available and thus 'accessible'. The gene expression and copy number data are in standard data matrix formats that support formal sharing and satisfy the 'interoperable' condition. Finally, this dataset can be 'reused' for additional research through either the G-DOC platform or GEO.

Commandment 9: Make sure that successors are being instructed correctly

Translational research projects usually take 4-5 years, which is a long period of time. Clinicians, researchers and data managers, but also trial nurses, might come and go. In the case these trial nurses performed the data entry for the study, they probably spent quite some time learning how to enter data into the eCRF. To avoid that the new data entry person needs to spend a similar amount of time to learn about this data entry, the old data entry person should properly instruct the new person, reducing the learning time. The same holds for the data managers. The leader of the data management WP (see commandment 1) might even make a data entry manual together with the data entry person, to ensure that any transfers of data entry tasks will go smoothly. As stated in commandment 2: data quality is extremely important and thus correct data entry should be a priority.

Commandment 10: Make it sustainable: what happens after the project?

When starting a new translational research project, big plans are made for the duration of the project, but very often not so much for the period after. What will happen when the project is finished? For example: who will pay for the continued storage of left-over biomaterials? Who will keep the database running? The researchers might even want to continue the project with yearly updates, because long-term follow-up information is actually really valuable in these type of projects. Or they want to submit the data to a repository such as Dataverse [49] or Dryad [50], if the informed consent allows it. Publicly available datasets can be a goldmine for future research [51], certainly with the rise of artificial intelligence methods. At the start of the project, the researchers should already make a plan for what happens at the end of the study, when funding runs out, to avoid that data and biomaterials are lost for future research. This planning should also include a financial paragraph, because hosting of data (and storage of biomaterials) will need to be paid for somehow, certainly if the data is not submitted to a public repository.

3. Summary and Conclusions

1	Create a separate Data Management work package
2	Reserve time and money for data entry
3	Define all data fields up front with the help of data analysis experts
4	Make clear arrangements about data access
5	Agree about de-identification and anonymization
6	Reuse existing software where possible
7	Make newly created software reusable
8	Adhere to the FAIR Guiding Principles
9	Make sure that successors are being instructed correctly
10	Make it sustainable: what happens after the project?

Table 2. Summary of the Ten Commandments of Translational Research Informatics

Translational research informatics is a field that is linked to data science and big data analytics, because of the ever growing size of the datasets and the need for analysis by machines. This means that the research output generated by the studies should be machine-readable, i.e. properly described by metadata, standardized according to ontologies, etc. [47]. The field is also heavily influenced by new privacy laws such as the GDPR: the infrastructure that is created needs to comply with stricter security and privacy rules than ever before. More emphasis is being placed on the importance of de-identification, pseudonymization and anonymization, certainly now that there is a trend to connect translational research informatics systems directly to the EHR [52], which contains personal data. Moreover, security measures such as multi-factor authentication (MFA) and data encryption are getting more common. The ten commandments presented in this article (see Table 2 for the summary) reflect the current state of the field, and might be subject change in a rapidly developing field. The rise of 'open science' and, related to this, the FAIR Guiding Principles, gives much-needed attention to data sharing, reuse of data and methods, reproducibility, etc. In some funding programs, such as Horizon 2020 from the EU, projects are already instructed to adhere to the FAIR Guiding Principles, and to create a Data Management Plan (DMP) which helps to think about data sharing, what will happen to the data after the project, etc. The other commandments listed here are mentioned less often in publications around data stewardship and data management, but are just as crucial for the success of a translational research project.

4. Competing interest statement

Dr. Hulsen is employed by Philips Research.

5. Disclaimer

This manuscript reflects an interpretation of the GDPR by the author, who is not a legal expert.

6. References

[1] P.R. Luijten, G.A. van Dongen, C.T. Moonen, G. Storm, and D.J. Crommelin, Public-private partnerships in translational medicine: concepts and practical examples, *J Control Release* **161** (2012), 416-421. PubMed ID: 22465390.

[2] D. Butler, Translational research: crossing the valley of death, *Nature* **453** (2008), 840-842. PubMed ID: 18548043.

[3] R. Becker and G.A. van Dongen, EATRIS, a vision for translational research in Europe, *J Cardiovasc Transl Res* **4** (2011), 231-237. PubMed ID: 21544739.

[4] C.P. Investigators, H. Shamoon, D. Center, P. Davis, M. Tuchman, H. Ginsberg, R. Califf, D. Stephens, T. Mellman, J. Verbalis, L. Nadler, A. Shekhar, D. Ford, R. Rizza, R. Shaker, K. Brady, B. Murphy, B. Cronstein, J. Hochman, P. Greenland, E. Orwoll, L. Sinoway, H. Greenberg, R. Jackson, B. Coller, E. Topol, L. Guay-Woodford, M. Runge, R. Clark, D. McClain, H. Selker, C. Lowery, S. Dubinett, L. Berglund, D. Cooper, G. Firestein, S.C. Johnston, J. Solway, J. Heubi, R. Sokol, D. Nelson, L. Tobacman, G. Rosenthal, L. Aaronson, R. Barohn, P. Kern, J. Sullivan, T. Shanley, B. Blazar, R. Larson, G. FitzGerald, S. Reis, T. Pearson, T. Buchanan, D. McPherson, A. Brasier, R. Toto, M. Disis, M. Drezner, G. Bernard, J. Clore, B. Evanoff, J. Imperato-McGinley, R. Sherwin, and J. Pulley, Preparedness of the CTSA's structural and scientific assets to support the mission of the National Center for Advancing Translational Sciences (NCATS), *Clin Transl Sci* **5** (2012), 121-129. PubMed ID: 22507116.

[5] P.R. Payne, S.B. Johnson, J.B. Starren, H.H. Tilson, and D. Dowdy, Breaking the translational barriers: the value of integrating biomedical informatics and translational research, *J Investig Med* **53** (2005), 192-200. PubMed ID: 15974245.

[6] T. Hulsen, S.S. Jamuar, A.R. Moody, J.H. Karnes, O. Varga, S. Hedensted, R. Spreafico, D.A. Hafler, and E.F. McKinney, From Big Data to Precision Medicine, *Frontiers in Medicine* **6** (2019).

[7] Research Data Management (A How-to Guide): Research Data Management Definition, https://libguides.depaul.edu/c.php?g=620925&p=4324498.

[8] A. Surkis and K. Read, Research data management, *J Med Libr Assoc* **103** (2015), 154-156. PubMed ID: 26213510.

[9] S. Rosenbaum, Data governance and stewardship: designing data stewardship entities and advancing data access, *Health Serv Res* **45** (2010), 1442-1455. PubMed ID: 21054365.

[10] Handbook for Adequate Natural Data Stewardship (HANDS) - Data Stewardship, https://data4lifesciences.nl/hands2/data-stewardship/.

[11] LIMA - Liquid Biopsies and Imaging, https://lima-project.eu/.

[12] RE-IMAGINE - Correcting Five Decades of Error through Enabling Image-based Risk Stratification of Localised Prostate Cancer, https://www.reimagine-pca.org/.

[13] T. Hulsen, J.H. Obbink, E.A.M. Schenk, M.F. Wildhagen, and C.H. Bangma, PCMM Biobank, IT-infrastructure and decision support, in, 2013.

[14] T. Hulsen, J.H. Obbink, W. Van der Linden, C. De Jonge, D. Nieboer, S.M. Bruinsma, R. M.J., and C.H. Bangma, 958 Integrating large datasets for the Movember Global Action Plan on active surveillance for low risk prostate cancer, *European Urology Supplements* **15** (2016), e958.

[15] T. Hulsen, W. Van der Linden, C. De Jonge, J. Hugosson, A. Auvinen, and M.J. Roobol, Developing a future-proof database for the European Randomized study

of Screening for Prostate Cancer (ERSPC), *European Urology Supplements* (2019). PubMed ID: 9088276.
[16] G.A. Meijer, J.W. Boiten, J.A.M. Beliën, H.M.W. Verheul, M.N. Cavelaars, A. Dekker, P. Lansberg, R.J.A. Fijneman, W. Van der Linden, R. Azevedo, and N. Stathonikos, TraIT - Translational Research IT, in, 2017.
[17] Horizon 2020 - The EU Framework Programme for Research and Innovation, https://ec.europa.eu/programmes/horizon2020/en.
[18] OpenClinica - Electronic Data Capture for Clinical Research, https://www.openclinica.com.
[19] Castor EDC - Cloud-based Electronic Data Capture Platform, https://www.castoredc.com.
[20] P.A. Harris, R. Taylor, R. Thielke, J. Payne, N. Gonzalez, and J.G. Conde, Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support, *J Biomed Inform* **42** (2009), 377-381. PubMed ID: 18929686.
[21] L. Cai and Y. Zhu, The Challenges of Data Quality and Data Quality Assessment in the Big Data Era, *Data Science Journal* **14** (2015).
[22] M.F. Kilkenny and K.M. Robinson, Data quality: "Garbage in - garbage out", *Health Inf Manag* **47** (2018), 103-105. PubMed ID: 29719995.
[23] T. Hulsen, W. Van der Linden, D. Pletea, J.H. Obbink, and M.J. Quist, Data Model Mapping, in, 2017.
[24] B. Smith and R.H. Scheuermann, Ontologies for clinical and translational research: Introduction, *J Biomed Inform* **44** (2011), 3-7. PubMed ID: 21241822.
[25] E.P.a.C.o.t.E. Union, Regulation on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (Data Protection Directive), *Official Journal of the European Union* **59** (2016), 1-88.
[26] U.S. Government, Health Insurance Portability and Accountability Act Of 1996, in, 1996.
[27] F. Prasser, F. Kohlmayer, R. Lautenschlager, and K.A. Kuhn, ARX--A Comprehensive Tool for Anonymizing Biomedical Data, *AMIA Annu Symp Proc* **2014** (2014), 984-993. PubMed ID: 25954407.
[28] DICOM Anonymizer, https://dicomanonymizer.com/.
[29] DicomCleaner, http://www.dclunie.com/pixelmed/software/webstart/DicomCleanerUsage.html.
[30] D.S. Marcus, T.R. Olsen, M. Ramaratnam, and R.L. Buckner, The Extensible Neuroimaging Archive Toolkit: an informatics platform for managing, exploring, and sharing neuroimaging data, *Neuroinformatics* **5** (2007), 11-34. PubMed ID: 17426351.
[31] E. Scheufele, D. Aronzon, R. Coopersmith, M.T. McDuffie, M. Kapoor, C.A. Uhrich, J.E. Avitabile, J. Liu, D. Housman, and M.B. Palchuk, tranSMART: An Open Source Knowledge Management and High Content Data Analytics Platform, *AMIA Jt Summits Transl Sci Proc* **2014** (2014), 96-101. PubMed ID: 25717408.
[32] J. Gao, B.A. Aksoy, U. Dogrusoz, G. Dresdner, B. Gross, S.O. Sumer, Y. Sun, A. Jacobsen, R. Sinha, E. Larsson, E. Cerami, C. Sander, and N. Schultz, Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal, *Sci Signal* **6** (2013), pl1. PubMed ID: 23550210.

[33] C. Costa, C. Ferreira, L. Bastiao, L. Ribeiro, A. Silva, and J.L. Oliveira, Dicoogle - an open source peer-to-peer PACS, *J Digit Imaging* **24** (2011), 848-856. PubMed ID: 20981467.
[34] E. Afgan, D. Baker, B. Batut, M. van den Beek, D. Bouvier, M. Cech, J. Chilton, D. Clements, N. Coraor, B.A. Gruning, A. Guerler, J. Hillman-Jackson, S. Hiltemann, V. Jalili, H. Rasche, N. Soranzo, J. Goecks, J. Taylor, A. Nekrutenko, and D. Blankenberg, The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update, *Nucleic Acids Res* **46** (2018), W537-W544. PubMed ID: 29790989.
[35] S.N. Murphy, G. Weber, M. Mendis, V. Gainer, H.C. Chueh, S. Churchill, and I. Kohane, Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2), *J Am Med Inform Assoc* **17** (2010), 124-130. PubMed ID: 20190053.
[36] Occhiolino - Laboratory Information Management System for Healthcare and Biomedicine, http://lims.gnu.org.
[37] L.D. McIntosh, M.K. Sharma, D. Mulvihill, S. Gupta, A. Juehne, B. George, S.B. Khot, A. Kaushal, M.A. Watson, and R. Nagarajan, caTissue Suite to OpenSpecimen: Developing an extensible, open source, web-based biobanking management system, *J Biomed Inform* **57** (2015), 456-464. PubMed ID: 26325296.
[38] S. Jodogne, The Orthanc Ecosystem for Medical Imaging, *J Digit Imaging* **31** (2018), 341-352. PubMed ID: 29725964.
[39] P. Bankhead, M.B. Loughrey, J.A. Fernandez, Y. Dombrowski, D.G. McArt, P.D. Dunne, S. McQuaid, R.T. Gray, L.J. Murray, H.G. Coleman, J.A. James, M. Salto-Tellez, and P.W. Hamilton, QuPath: Open source software for digital pathology image analysis, *Sci Rep* **7** (2017), 16878. PubMed ID: 29203879.
[40] SlideAtlas - Whole Slide Image Viewer, https://slide-atlas.org/.
[41] J. Perkel, Democratic databases: science on GitHub, *Nature* **538** (2016), 127-128. PubMed ID: 27708327.
[42] SourceForge - The Complete Open-Source and Business Software Platform, https://sourceforge.net/.
[43] J. Singh, FigShare, *J Pharmacol Pharmacother* **2** (2011), 138-139. PubMed ID: 21772785.
[44] Zenodo - Research. Shared., https://zenodo.org/.
[45] P.C. Griffin, J. Khadake, K.S. LeMay, S.E. Lewis, S. Orchard, A. Pask, B. Pope, U. Roessner, K. Russell, T. Seemann, A. Treloar, S. Tyagi, J.H. Christiansen, S. Dayalan, S. Gladman, S.B. Hangartner, H.L. Hayden, W.W.H. Ho, G. Keeble-Gagnere, P.K. Korhonen, P. Neish, P.R. Prestes, M.F. Richardson, N.S. Watson-Haigh, K.L. Wyres, N.D. Young, and M.V. Schneider, Best practice data life cycle approaches for the life sciences, *F1000Res* **6** (2017), 1618. PubMed ID: 30109017.
[46] P.H. Russell, R.L. Johnson, S. Ananthan, B. Harnke, and N.E. Carlson, A large-scale analysis of bioinformatics code on GitHub, *PLoS One* **13** (2018), e0205898. PubMed ID: 30379882.
[47] M.D. Wilkinson, M. Dumontier, I.J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.W. Boiten, L.B. da Silva Santos, P.E. Bourne, J. Bouwman, A.J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C.T. Evelo, R. Finkers, A. Gonzalez-Beltran, A.J. Gray, P. Groth, C. Goble, J.S. Grethe, J. Heringa, P.A. t Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S.J. Lusher, M.E. Martone, A. Mons, A.L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M.A. Swertz, M. Thompson, J. van der Lei, E. van

Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, and B. Mons, The FAIR Guiding Principles for scientific data management and stewardship, *Sci Data* **3** (2016), 160018. PubMed ID: 26978244.
[48] Y. Gusev, K. Bhuvaneshwar, L. Song, J.C. Zenklusen, H. Fine, and S. Madhavan, The REMBRANDT study, a large collection of genomic data from brain cancer patients, *Sci Data* **5** (2018), 180158. PubMed ID: 30106394.
[49] B. McKinney, P.A. Meyer, M. Crosas, and P. Sliz, Extension of research data repository system to support direct compute access to biomedical datasets: enhancing Dataverse to support large datasets, *Ann N Y Acad Sci* **1387** (2017), 95-104. PubMed ID: 27862010.
[50] H.C. White, S. Carrier, A. Thompson, J. Greenberg, and R. Scherle, The Dryad data repository: a Singapore framework metadata architecture in a DSpace environment, *DCMI '08 Proceedings of the 2008 International Conference on Dublin Core and Metadata Applications* (2008), 157-162.
[51] T. Hulsen, An overview of publicly available patient-centered prostate cancer datasets, *Transl Androl Urol* (2019).
[52] Y.L. Yip, Unlocking the potential of electronic health records for translational research. Findings from the section on bioinformatics and translational informatics, *Yearb Med Inform* **7** (2012), 135-138. PubMed ID: 22890355.