**The Ten Commandments of Translational Research Informatics**

Tim HULSENa,1

aDepartment of Professional Health Solutions & Services, Philips Research, Eindhoven, The Netherlands

1Corresponding author. E-mail: [tim.hulsen@philips.com](mailto: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 productized. 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. The author has worked on the IT infrastructure for several translational research projects in oncology for the past nine years, and presents his learnings in this paper in the form of ten commandments. These learnings 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, which might mean that they no longer need to be mentioned in an overview like this within a few years. The other commandments might still be noteworthy in many years to come.

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

# 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 productized. 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]. The author has worked on the IT infrastructure, data integration and data management for several translational research projects in oncology [7-11] for the past nine years, as well as the Dutch translational research informatics project CTMM-TraIT [12], and presents his learnings in this paper in the form of ten commandments. These learnings 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.

# The Ten Commandments

## Commandment 1: Create a separate Data Management WP

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 institutes, and nobody knows how to combine them and make sense of them. The solution: create a separate work package on data management, which 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, 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 data into an EDC system such as OpenClinica [13] or Castor EDC [14]. 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 [15]. The term “Garbage in, garbage out” (GIGO) comes from computer science, but applies to medical data as well [16]. 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.

## Commandment 3: Define all data fields upfront

Within PCMM [9], we noticed after a few years that some information that was essential to answer certain research questions was not being collected in the 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 [10], 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. Besides the data field name, we stored the data type (integer, float, string, categorical) and the unit (cm, kg, ng/ml, etc.), and (in case of a categorical data type) listed the categories. Because part of the Movember GAP3 data was retrospective, we created some data model mapping scripts [17] to map the existing data to this codebook.

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 BFO, OBO and RO [18].



**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 upfront. Of course, data access needs to be arranged in the informed consent as well, as patients are the data owners, and the GPDR [19] (EU) and HIPAA [20] (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.

At the end of the project, data should be shared with the whole world. 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.

## Commandment 5: Agree about de-identification and anonymization

The responsibility for proper de-identification of the data lies with the organization that collects the data (usually the hospital), because they are the ones that have the 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, it should be destroyed immediately because of the privacy risk. 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. For textual and numerical data, open-source software packages are available that can help with anonymization, such as the ARX anonymization tool [21]. For imaging data, anonymization tools are available that can strip any identifiable information from DICOM tags, and can even edit the image itself [22].

## 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 TraIT project [12], a list of suitable open source tools was created, which included OpenClinica [13], XNAT [23] and tranSMART [24]. 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.

|  |  |  |
| --- | --- | --- |
| **Name** | **Description** | **URL** |
| cBioPortal [25] | Open source tool for cancer genomics. | <https://github.com/cBioPortal/> |
| Dicoogle [26] | Open source Picture Archiving and Communications System (PACS) archive. | <http://www.dicoogle.com/> |
| Galaxy [27] | Open source, web-based platform for data intensive biomedical research. | <https://usegalaxy.org/> |
| I2B2 [28] | Informatics for Integrating Biology & the Bedside. | <https://www.i2b2.org/> |
| Occhiolino [29] | GNU LIMS, also known as Occhiolino, is a modern, scalable Laboratory Information Management System for the healthcare and biomedical sectors. | <http://lims.gnu.org/> |
| OpenClinica Community Edition [13] | 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 [30] | Open source informatics platform for biobanking. | <https://www.openspecimen.org/> |
| Orthanc [31] | Open source, lightweight DICOM server. | <https://www.orthanc-server.com/> |
| QuPath [32] | Open source software for Quantitative Pathology. | <https://qupath.github.io/> |
| REDCap [33] | Research Electronic Data Capture. | <https://redcap.ahc.umn.edu/> |
| SlideAtlas [34] | Open-source, web-based, whole slide imaging platform. | https://slide-atlas.org/ |
| tranSMART [24] | Open source suite of data exploration, visualization, and ETL tools, developed for translational research studies. | <https://transmartfoundation.org/current-transmart-platform-release/> |
| XNAT [23] | Open source imaging informatics platform. | <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 [35], SourceForge [36] or FigShare [37], or it can be made available on a custom-made website, and then referenced on Zenodo [38]. Griffin et al. [39] 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 [40].

## Commandment 8: Adhere to the FAIR Guiding Principles

In 2016, the FAIR Guiding Principles for scientific data management and stewardship [41] 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.

## Commandment 9: Make sure that any replacements 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: Think about sustainability: 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 [42] or Dryad [43], if the informed consent allows it. Publicly available datasets can be a goldmine for future research [44], 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.

# Summary and Conclusions

**Figure 3.** Stone tablet with 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. [41]. 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 [45], 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 Figure 3 for the summary) reflect the current state of the field, and are actually not really written on a stone tablet: in the (near) future it might be that some of the commandments are already obsolete, simply because everybody adheres to them already. 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. So, hopefully, in five years a manuscript like this will be obsolete, or at least significantly shorter.

# Competing interest statement

Dr. Hulsen is employed by Philips Research.

# Disclaimer

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

# 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] LIMA - Liquid Biopsies and Imaging, in.

[8] RE-IMAGINE - Correcting Five Decades of Error through Enabling Image-based Risk Stratification of Localised Prostate Cancer, in.

[9] C.H. Bangma, E. Schenk, G. Jenster, I.J. de Jong, H. Obbink, T. Hulsen, and E. Caldenhoven, Prostate Cancer Molecular Medicine - Molecular diagnosis, prognosis and imaging for tailored treatment of prostate cancer, in, 2016.

[10] 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.

[11] 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.

[12] 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.

[13] OpenClinica - Electronic Data Capture for Clinical Research, in.

[14] Castor EDC - Cloud-based Electronic Data Capture Platform, in.

[15] L. Cai and Y. Zhu, The Challenges of Data Quality and Data Quality Assessment in the Big Data Era, *Data Science Journal* **14** (2015).

[16] M.F. Kilkenny and K.M. Robinson, Data quality: "Garbage in - garbage out", *Health Inf Manag* **47** (2018), 103-105. PubMed ID: 29719995.

[17] T. Hulsen, W. Van der Linden, D. Pletea, J.H. Obbink, and M.J. Quist, Data Model Mapping, in, 2017.

[18] B. Smith and R.H. Scheuermann, Ontologies for clinical and translational research: Introduction, *J Biomed Inform* **44** (2011), 3-7. PubMed ID: 21241822.

[19] 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.

[20] U.S. Government, Health Insurance Portability and Accountability Act Of 1996, in, 1996.

[21] 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.

[22] E. Monteiro, C. Costa, and J.L. Oliveira, A De-Identification Pipeline for Ultrasound Medical Images in DICOM Format, *J Med Syst* **41** (2017), 89. PubMed ID: 28405948.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] Occhiolino - Laboratory Information Management System for Healthcare and Biomedicine, in.

[30] 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.

[31] S. Jodogne, The Orthanc Ecosystem for Medical Imaging, *J Digit Imaging* **31** (2018), 341-352. PubMed ID: 29725964.

[32] 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.

[33] 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.

[34] SlideAtlas - Whole Slide Image Viewer, in.

[35] J. Perkel, Democratic databases: science on GitHub, *Nature* **538** (2016), 127-128. PubMed ID: 27708327.

[36] SourceForge - The Complete Open-Source and Business Software Platform, in.

[37] J. Singh, FigShare, *J Pharmacol Pharmacother* **2** (2011), 138-139. PubMed ID: 21772785.

[38] Zenodo - Research. Shared., in.

[39] 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.

[40] 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.

[41] 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.

[42] 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.

[43] 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.

[44] T. Hulsen, An overview of publicly available patient-centered prostate cancer datasets, *Transl Androl Urol* (2019).

[45] 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.