Health Informatics Meets eHealth
D. Hayn and G. Schreier (Eds.)

doi:10.3233/978-1-61499-759-7-70

Two Years of tranSMART in a University Hospital for Translational Research and Education

Jan CHRISTOPH[a,1], Christian KNELL[a], Elisabeth NASCHBERGER[b], Michael STÜRZL[b], Christian MAIER[a], Hans-Ulrich PROKOSCH[a] and Martin SEDLMAYR[a]
[a]*Chair of Medical Informatics, Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany*
[b]*Division of Molecular and Experimental Surgery, Department of Surgery, Translational Research Center Erlangen, Universitätsklinikum Erlangen, Germany*

Abstract. *Background*: For translational research, software platforms such as tranSMART with an integrated view of both clinical and omics data have gained more and more attention in the last years. *Objectives*: We wanted to examine the success and failures of tranSMART in the fields of translational research and education by looking at the examples of six use cases at our hospital. We wanted to point out suitable scenarios and user groups as well as still existing challenges and limitations. *Methods*: We sum up the experience we made with our use cases with a focus on lessons learned. *Results*: tranSMART was successfully established by a bottom-up approach at our university hospital and has been running for more than two years now. It has been used in four translational research projects as well as in education in the context of lectures and bachelor/master theses. *Conclusion*: tranSMART can be a very useful tool for translational research and education. But it should be used with both care and statistical knowledge to avoid wrong conclusions. Some technical constraints, especially for data modeling, still limit many applications. Version control and data provenance are remaining challenges.

Keywords. Translational Medical Research, Biomedical Research, Research Personnel, Medical Education.

1. Introduction

1.1. Background and related work

Translational research requires the integration of heterogeneous data (both clinical and omics data) into a unified view for analysis [1]. Such a view can provide researchers with a source of both, generation and validation of hypotheses and of cohort selection and biomarker discovery. It might also enable the reuse of valuable existing data with the consequence of a reduction of costs and an increase in research effectiveness [2]. Like for Bauer et al. [3], *integration* means for us that "1) the different data types of interest are accessible via a single platform, 2) that data are cross-referenced (i.e. different types of patient-specific data such as molecular and clinical data can be linked) and 3) that data formats and platform infrastructures facilitate querying".

[1] Corresponding Author: Jan Christoph, Chair of Medical Informatics, Friedrich-Alexander-Universität Erlangen-Nürnberg, Wetterkreuz 13, 91058 Erlangen, Germany. E-Mail: jan.christoph@fau.de

Tools like cBioPortal [4], iDASH [5], and tranSMART [6] provide such an integrated view. They enable the efficient exploration of data by presenting both clinical and omics data in an integrated way. They also provide already existing functions for data exploration, cohort selection and the generation and validation of hypotheses.

The tranSMART platform has its roots in the i2b2 phenotype framework [7]. An active open-source community organized by the tranSMART Foundation has been developing it since 2013. In its database, tranSMART combines clinical data in an entity-attribute-value store like i2b2 [7] with separate tables for omics data. On top, an application server provides the user with an intuitive web front-end.

There are already reviews comparing these platforms [8,9] and publications exist which deal with tranSMART and mention its use for clinical studies, e.g. [10]. Furthermore, there are papers which picture tranSMART in detail and from a technical point of view: They describe the platform and the architecture of the system [6,11], its meaning in the context of a study infrastructure [2,3,12], and in connection with other platforms (like Galaxy, Minerva or Genedata Analyst) [13,14] or technologies (like NoSQL or HL7) [15,16].

1.2. Motivation and Objectives

At our university hospital, high-throughput omics technologies [17] have become an integral part of medical research, especially in oncology. Interdisciplinary teams of biometricians, molecular biologists and physicians analyze traditional clinical data as well as omics data in considerable amount and across different levels, e.g. single-nucleotide polymorphisms (SNPs) by the genome, gene expressions by the transcriptome or qualitative signals by the proteome. Simultaneously, our department of medical informatics has been asked with increasing frequency to support these teams with standardized services for storing, querying and analyzing such data. The researchers expressed their need for 1) secondary use of data from routine systems [18], 2) integration of both clinical and omics data into a comprehensive data warehouse, 3) graphical and easy-to-use tools for analyzing such data, 4) collaboration tools for exchanging knowledge and data between biometricians and physicians and 5) training of students and experts.

We identified tranSMART as a tool which is capable of solving these requirements and began to establish it in September 2014 as service to our researchers. Since then, it has used in four translational research projects as well as in education (lectures and bachelor/master theses of medical students and computer scientists).

To the best of our knowledge, there is no publication about tranSMART yet which 1) outlines the adoption of this platform - not as a tool in the specific research project but as a bottom-up approach for the general support of translational research groups - and 2) describes the use of tranSMART for educational purposes. In the following, we want to describe the types of projects in which tranSMART was used, report on our observations and experiences, and discuss some limitations we have encountered.

2. Methods

2.1. Getting started: Identifying suitable research partners and platform

In June 2014, we ranked all institutes (>100 units) of our university hospital with regard to their clinical omics-activities based on the annual research report and literature research. The leaders of the top-ten-ranked research groups were asked for an interview, which was accepted in nine cases. Each took place in a semi-structured manner in two parts: The first part comprised questions on running and planned projects, used methods, resulted or expected data and finally the noticed limitations for a more efficient research. In the second part of the interview tranSMART was demonstrated as an example of an integration platform with public data to determine whether this kind of software is considered to be useful in principle or not. Three suitable partners (having both need and data and covering a broad range of use cases) were selected to establish a prototype of an integrating platform for translational research. The fourth use case arose shortly afterwards in the form of a big research project with several external partners who needed a research database for the integration of both clinical and omics data [19].

For the final determination of the platform, the review of Canuel et al. [9] was a very useful starting point to select tranSMART as a platform which fulfills our requirements, which, among others, arose by the interviews described above.

2.2. Installation, Data modeling and data loading

At the beginning, tranSMART 1.1 was installed, later we migrated to every new version immediately after its release - until the latest version 16.2. Our installations are based on Ubuntu Linux 14.04 and use PostgreSQL 9.3 as a database in virtual machines (VM) at an ESX-cluster. There are three tranSMART-VMs for development, quality assurance and productive mode within our internal hospital network. The forth VM is publicly available and used for demonstration and education purposes. All VMs have 8 GB RAM, 4 CPUs and between 50 GB and 500 GB hard disk.

In all four use cases of translational research, we modeled the data by first using Microsoft Excel files in an iterative process between physicians and computer scientists. Finally, the extract-transform-loading (ETL-)-Tool Talend Open Studio Data Integration performed the transformation of the raw data into data which were ready for being uploaded to tranSMART.

Since there are several tools to finally upload clinical data and different types of omics data into the tranSMART database, we generate a systematic overview and performed practical testing of all available tools. In the end, tMDataLoader [20] was chosen because it supports most data types as shown in Table 1.

2.3. Using tranSMART for education and spreading it to further researchers

Since it is open source, tranSMART offers many possibilities for extensions and applications. That is why we announced topics for bachelor or master theses for students of computer science (educational use case 1). Furthermore, we integrated tranSMART into lectures and used it for practical exercises (educational use case 2). Training material in the form of slides and screen videos has been prepared and an exercise with 18 tasks has been designed to give an overview of the handling of tranSMART and its most

Table 1: Overview of all public available ETL-tools at September, 2016. The column *Support* indicates the general support of the datatype. Column *HDD* shows whether an import of the datatype as high-dimensional data is possible. *XX*: import is officially supported by the tool and could be reproduced by us. *XN*: our try of import was not successful. *X?*: we weren't able to test the import due to missing data. *NN*: import is officially not supported by the tool. ---: Due to technical constraints by tranSMART, import is technically not possible.

Datatype / ETL-tool	**transmart-data**		**tMDataLoader**		**transmart-batch**		**ICE (Int. Cur. Environment)**	
	Support	**HDD**	**Support**	**HDD**	**Support**	**HDD**	**Support**	**HDD**
Clinical	XX	---	XX	---	XX	---	XX	---
aCGH / CNV	XX	XX	XX	XX	XX	XX	NN	NN
cDNA	X?	---	X?	---	X?	---	X?	---
Methylation	X?	---	X?	---	X?	---	X?	---
miRNA	XX	XX	XX	XX	XX	XX	XX	XX
mRNA	XX	XX	XX	XX	XX	XX	XX	XX
Proteomics	XX	XN	XX	XX	XX	XX	XX	XX
RNASeq	XX	XX	XX	XX	XX	XX	XX	XX
SNP	NN	NN	XX	XX	NN	NN	XX	XX
VCF	XX	XX	XX	XX	NN	NN	NN	NN

important functions. Meanwhile, it has become apparent to us that many of our translational researchers would use tranSMART, but are not aware of its existence or have no experience with it. Therefore, an adapted version of the student exercise was mailed to 45 medical-focused researchers (identified amongst others by the rating described above) and 27 of them spent about an hour with it and gave mostly very positive feedback with the request to use tranSMART also for their own data and research.

2.4. Evaluation and Logging

For the last two years, we have collected every tranSMART-related feedback that we received from researchers and students. For the first three uses cases of translational research, we conducted a semi-structured interview with the researchers six weeks after their use of tranSMART. Furthermore, we evaluated the access logs provided by the admin interface of tranSMART as well as the Tomcat logs by an Elasticsearch, Logstash, and Kibana (ELK) stack [21].

3. Result

In the following, the main use cases are described which have been supported by tranSMART at our university hospital.

3.1. Research use case 1: Supporting the analysis of an ongoing prospective study of colorectal cancer of the Division of Molecular and Experimental Surgery.

The final dataset of more than 600 patients was derived from a CSV-export of an electronic data capture system which provided over 500 clinical items with six electronic case report forms (eCRF) per patient (baseline and up to five follow-ups). Furthermore, we integrated two Excel files with about 50 gene expressions (generated by a RT-qPCRanalysis) and two protein quantifications, respectively for each patient. About 15 additional items had to be calculated within the ETL-process since they were not explicitly available within the raw data (like *survival time* or *disease-free survival*) or their original categorization had to be transformed to be suitable for analyses (e.g. tumor locations in only three categories than in twelve as defined in the eCRF). To support querying and data exploration in tranSMART before version 16.2, gene expression data were modeled both as clinical and as high dimensional data. This dataset was complemented by clinical data (13 items) and gene expression data (120 genes) of 177 patients of a comparable public study (GEO GSE17536 [22]) to support the generation of hypotheses. Every six months, the whole ETL-process is repeated to add new data which has been gained from further follow-ups and further omics-analyses. The whole iterative process of data modeling and the establishment of the periodic ETL took about 500-750h for us. In return, the resulting tranSMART-project has been intensively analyzed by four researchers of the department with altogether about 1.000 logins during the last two years and with an average time of use of about one hour per login.

3.2. Research use case 2: Supporting the final analysis of a completed retrospective study of cancer vaccination of the Department of Dermatology.

The raw data of 62 patients - with each patient having 37 clinical items of up to 42 visits - was provided as an Excel file. The data contained no omics data and had to be imported only once. Since tranSMART still lacks a feasible “time-series” concept, each parameter had to be summarized over all visits of a patient in one single value. In case of numerical variables, we used the minimum, the maximum or the average value. Categorical variables were discussed with a physician to determine the mapping rule on how to represent different categorical values over all visits in a final value. Modeling and importing the data took about 40 hours for us. The resulting tranSMART-project was analyzed by two users for about three months with altogether about 50 logins with an average time of use of about 1,5h per login.

3.3. Research use case 3: Supporting the analysis of a completed retrospective study on gastrointestinal stroma tumors of the department of Internal Medicine.

The rather small dataset of four clinical items (no omics data) of 32 patients was available in the form of an Excel file. It could be imported straight forward within half an hour. The resulting tranSMART-project was analyzed in the context of a PhD thesis of a medical student who performed survival analyses for one week with altogether six logins with an average time of use of about 30 minutes per login.

3.4. Research use case 4: Providing a central research database in an interdisciplinary consortium of more than five partners.

This use case differs from the first three use cases as the main focus has been on the data integration and enrichment of the data rather than on data analysis. The initial dataset contained about 300 clinical data items as well as omics data (about 18.000 gene expressions and almost 200.000 SNPs) of 812 patients of a completed clinical study on breast cancer. In the course of the project, this data body was added by deidentified free-text (e.g. from discharge letters) as well as structured data from text mining and imaging analysis. Data modeling was performed over an extended period of two years - depending on the availability of data in the research project.

While the data has been successfully imported, the resulting tranSMART-project unfortunately has not been used as a source of information. It turned out that the use of relational-structured source files (Excel, CSV) directly seemed to be more straightforward to programmers. No added value was perceived in using the entity attribute-value scheme or the RESTful-API provided by tranSMART. Researchers who wanted to develop new analysis methods based on these data stated that they would rather have preferred the use of a common data model such as OMOP/OHDSI [23].

Across these use cases, the five analysis methods of tranSMART which have been used most often via the web interface are in descended order *summary statistics*, *ANOVA*, *survival analysis, fisher-test,* and *correlation analysis*. The feedback of researchers who used the analysis methods of tranSMART via the web front-end was very positive (e.g. "*tranSMART opens a magnitude of new opportunities for our future research!*", "*Also yesterday, I worked with great delight with this program.", "It is a dream.", "Greetings from the dermatology casino: I'm very pleased with my new toy!"*).

3.5. Educational use case 1

For two years, tranSMART has been the study object of various bachelor and master theses and of internships at our institution. As part of such a work, tranSMART has been established at our university hospital as a research service. We reconstructed the use of HBase for omics data as described by Wang et al. [16] and implemented an own workflow for interactive analysis by SmartR. In ongoing projects, we are trying to incorporate imaging data by using XNAT as suggested by He et al. [24] and we are attempting to connect tranSMART to the SMART-on-FHIR platform similar to the i2b2 example [25].

3.6. Educational use case 2

tranSMART has been used successfully in lectures as an example of IT-support for translational research. An exercise was prepared for using the web front-end of tranSMART for data exploration, cohort identification and for the generation and validation of hypotheses. The exercise included 18 tasks and took about an hour.

During the winter semester 2016, more than 120 medical students of the fifth semester performed this exercise. On average, they solved 90% of the tasks correctly, mostly having difficulties to validate hypotheses. The feedback provided by the students was very positive: more than 40 times they stated explicitly that "*it was fun*", "*it makes sense for my academic studies*", "*by the exercise I got the desire for a statistical dissertation*", "*this exercise showed that it is not too difficult to analyze data with the right tool*" and the like.

4. Discussion and lessons learned

4.1. User profiles and scenarios

The overall perception of tranSMART among researchers as well as students was very positive. It is considered a useful tool for translational research which has the ability to integrate clinical and omics data which provides many methods for analyzing such data.

However, having omics data to analyze does not automatically guarantee success if researchers already have toolchains in place which are not compatible with tranSMART (use case 4). On the other hand, tranSMART can be used successfully even without omics data (use case 2 and 3). The size of the study (number of data elements) seems to be less important (use case 3) as long as it justifies the data modeling and import efforts which can take in the shortest time 30 minutes from scratch until several months.

TranSMART is especially suited for cohort identification, data exploration, and the generation and validation of hypotheses. The user can choose predefined functions such as correlation analysis, ANOVA or survival analysis from a catalog and parametrize them according to his/her needs. Although it is possible to build own methods within the SmartR-workflows, it will likely not be successful to perform basic research in the field of computational molecular biology with tranSMART. The reason of this downside is that the platform - despite its RESTful-API – is not designed to integrate programming code like R or Python to solve a chain of small individual problems with intermediate results.

Most medical-focused researchers who were interested in the use of tranSMART had average computer skills but were not experienced with programming languages like R and had no more than basic knowledge of statistic programs like SPSS. Those who were more experienced considered tranSMART to be more useful for education than for their own research.

Having an easy-to-use interface is also a pitfall and makes it essential to have some statistical background and the awareness that tranSMART does not check whether even basic preconditions for statistical tests are fulfilled (e.g. a sufficient sample size and a normal distribution for the t-test). Although these could still be checked automatically in theory, the user still has to be aware of confounder etc. and take them into consideration. This aspect can be – and has been - conveyed in a playful and interactive manner and it has, in fact, been highly appreciated by the students.

4.2. Data modeling, ETL and data loading

Data modeling is a crucial step prior to the import of data because it is neither possible to correct values within tranSMART nor to amend attributes. For example, if the difference between the *survival time* and the *disease-free survival time* is required for analysis, the difference has to be calculated and imported as an additional attribute already during the ETL process. Pseudonymization is also not supported - the patient identifier is visible via the feature *grid view* in the exact same manner as it has been imported.

The model of the clinical data requires to have exactly one line per patient with the same superset of variables as columns for all patients. This makes it challenging to represent 1:n relationships between the data such as multiple diagnoses of a patient or time series of lab values. Luckily, in our use cases, we were able to work around this limitation by aggregation and mapping rules, but it depends on each use case if a satisfying solution can be found. A major limitation of the data model is a lack of relationships between attributes, such as primary and secondary diagnosis or dependent attributes describing a finding: according to the tranSMART Foundation, the next release 17.1 – expected in summer 2017 - is to remove this limitation by becoming more compatible with i2b2 which uses therefor the concept of *modifiers*.

The tool tMDataLoader fulfilled all our current needs to import the data after the modeling process. However, in regard of clinical data it shows the same lack of incremental data updates and the limitation to flat CSV files (no XML or ODM) as all other publicly available tools.

4.3. Technic

The tranSMART-application has run sufficiently stable on our servers except for some sporadic overutilization of the CPU which rendered the web interface unusable for researchers.

The documentation provided by the community is more extensive than e.g. for i2b2 but lacks details especially for plugins like SmartR.

Although it should be possible to create and modify plugins, we found it quite arduous: the development of an own SmartR-workflow for survival analysis took several months. Moreover, tranSMART does not yet support distributed computing so that we performed our GWAS calculation externally on SparkR in a Hadoop-cluster [26].

4.4. Validation, reproducibility, archiving and data provenance

After importing the data into tranSMART, it is highly recommended to validate the data; A biometrician, for example, might use the R-interface to access the project and the exactly same data which have been analyzed by the physician before by the web front-end.

Even if the reproducibility of the analysis results is important for researchers, it cannot be guaranteed throughout tranSMART versions. For example, major changes occurred in the results of the survival analysis between version 1.2.3 and 1.2.4. due to minor changes in the R-scrips. Although only one character had been changed in the source code[2], the resulting negation yields completely different results of the analysis. Unfortunately, there was no notification or documentation about this change in the programmer's change log or elsewhere. Even if the documentation of tranSMART was complete, the checks and the validation of the platform would be hampered by the dependencies on numerous external libraries such as R-packages. An updated R-package, for example, caused missing values in a survival-related table after the tranSMART upgrade from version 1.2.5 to 16.1[3].

As a consequence, it is necessary to validate the imported data as well as the software configuration including all dependencies. Snapshots of the virtual machine are used to secure setups; but as this procedure is quite storage intensive, we consider using a more lightweight technology such as Docker in the future. Additionally, it would be beneficial, if tranSMART would support reproducibility by version management similar to other tools such as Galaxy.

5. Conclusion

As translational research becomes more important, an increasing number of medical-oriented researchers needs to integrate clinical and omics data for analysis. tranSMART has been proven as a viable platform which provides many required tools. In all four of

[2] *status <- currentDataSubset[[censor.field]]* changed to *status <- !currentDataSubset[[censor.field]]*

[3] Version 16.1 is the direct successor of 1.2.5 due to a new name convention.

our use cases, the data could be modeled and imported although the platform was only a final success in three cases.

Data modeling was the most important task in the provision of tranSMART as a service and requires an in-depth knowledge of the domain which is probably more crucial than the choice of the platform. While tranSMART provides functions for data analysis in an intuitive and graphical way it still requires basic statistical knowledge as it does not exempt the user from testing the prerequisites of an analysis. Using tranSMART for education as part of lectures and exercises for medical students and for computer scientists helps to sensitize future users.

TranSMART has been appreciated at our university hospital by translational researchers (*"tranSMART opens a magnitude of new opportunities for our future research!"*) and will be fostered further as a productive service for interested groups.

Acknowledgments

The research has been supported by the Smart Data Program of the German Federal Ministry for Economic Affairs and Energy (1MT14001B). The present work was performed in fulfillment of the requirements for obtaining the degree "Dr. rer. biol. hum." from the Friedrich-Alexander-Universität Erlangen-Nürnberg (JC). We thank Mrs. B. Schuler-Thurner (Department of Dermatology, Friedrich-Alexander Universität Erlangen-Nürnberg) for her cooperation and participation to adopt tranSMART at her department in research use case 2. Finally, we thank Stephanie Newe for linguistically proof-reading.

6. References

[1] Halevy, A., Rajaraman, A., and Ordille, J. 2006. Data integration: the teenage years. In Proceedings of the 32nd international conference on Very large data bases, 9–16.

[2] Jonnagaddala, J., Croucher, J. L., Jue, T. R., Meagher, N. S., Caruso, L., Ward, R., and Hawkins, N. J. 2016. Integration and Analysis of Heterogeneous Colorectal Cancer Data for Translational Research. Studies in health technology and informatics 225, 387–391.

[3] Bauer, C. R., Knecht, C., Fretter, C., Baum, B., Jendrossek, S., Ruhlemann, M., Heinsen, F.-A., Umbach, N., Grimbacher, B., Franke, A., Lieb, W., Krawczak, M., Hutt, M.-T., and Sax, U. 2016. Interdisciplinary approach towards a systems medicine toolbox using the example of inflammatory diseases. Briefings in bioinformatics.

[4] Gao, J., Aksoy, B. A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S. O., Sun, Y., Jacobsen, A., Sinha, R., Larsson, E., Cerami, E., Sander, C., and Schultz, N. 2013. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Science signaling 6, 269, pl1.

[5] Ohno-Machado, L., Bafna, V., Boxwala, A. A., Chapman, B. E., Chapman, W. W., Chaudhuri, K., Day, M. E., Farcas, C., Heintzman, N. D., Jiang, X., Kim, H., Kim, J., Matheny, M. E., Resnic, F. S., and Vinterbo, S. A. 2012. iDASH: integrating data for analysis, anonymization, and sharing. Journal of the American Medical Informatics Association : JAMIA 19, 2, 196–201.

[6] Athey, B. D., Braxenthaler, M., Haas, M., and Guo, Y. 2013. tranSMART: An Open Source and Community-Driven Informatics and Data Sharing Platform for Clinical and Translational Research. AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science 2013, 6–8.

[7] Kohane, I. S., Churchill, S. E., and Murphy, S. N. 2012. A translational engine at the national scale: informatics for integrating biology and the bedside. Journal of the American Medical Informatics Association : JAMIA 19, 2, 181–185.

[8] Dunn, W., JR, Burgun, A., Krebs, M.-O., and Rance, B. 2016. Exploring and visualizing multidimensional data in translational research platforms. Briefings in bioinformatics.

[9] Canuel, V., Rance, B., Avillach, P., Degoulet, P., and Burgun, A. 2015. Translational research latforms integrating clinical and omics data: a review of publicly available solutions. Briefings in bioinformatics 16, 2, 280–290.

[10] Haas, M., Stephenson, D., Romero, K., Gordon, M. F., Zach, N., and Geerts, H. 2016. Big data to smart data in Alzheimer's disease: Real-world examples of advanced modeling and simulation. Alzheimer's & dementia : the journal of the Alzheimer's Association 12, 9, 1022–1030.

[11] Scheufele, E., Aronzon, D., Coopersmith, R., McDuffie, M. T., Kapoor, M., Uhrich, C. A., Avitabile, J. E., Liu, J., Housman, D., and Palchuk, M. B. 2014. tranSMART: An Open Source Knowledge Management and High Content Data Analytics Platform. AMIA Summits on Translational Science Proceedings 2014, 96–101.

[12] Rance, B., Canuel, V., Countouris, H., Laurent-Puig, P., and Burgun, A. 2016. Integrating Heterogeneous Biomedical Data for Cancer Research: the CARPEM infrastructure. Applied clinical informatics 7, 2, 260–274.

[13] Satagopam, V., Gu, W., Eifes, S., Gawron, P., Ostaszewski, M., Gebel, S., Barbosa-Silva, A., Balling, R., and Schneider, R. 2016. Integration and Visualization of Translational Medicine Data for Better Understanding of Human Diseases. Big data 4, 2, 97–108.

[14] Schumacher, A., Rujan, T., and Hoefkens, J. 2014. A collaborative approach to develop a multi-omics data analytics platform for translational research. Applied & translational genomics 3, 4, 105–108.

[15] Camacho Rodriguez, J. C., Staubert, S., and Lobe, M. 2016. Automated Import of Clinical Data from HL7 Messages into OpenClinica and tranSMART Using Mirth Connect. Studies in health technology and informatics 228, 317–321.

[16] Wang, S., Pandis, I., Wu, C., He, S., Johnson, D., Emam, I., Guitton, F., and Guo, Y. 2014. High dimensional biological data retrieval optimization with NoSQL technology. BMC genomics 15 Suppl 8, S3.

[17] Thomas, D. C. 2006. High-volume "-omics" technologies and the future of molecular epidemiology. Epidemiology (Cambridge, Mass.) 17, 5, 490–491.

[18] Prokosch, H. U. and Ganslandt, T. 2009. Perspectives for medical informatics. Reusing the electronic medical record for clinical research. Methods of information in medicine 48, 1, 38–44.

[19] Sonntag, D., Tresp, V., Zillner, S., Cavallaro, A., Hammon, M., Reis, A., Fasching, P. A., Sedlmayr, M., Ganslandt, T., Prokosch, H.-U., Budde, K., Schmidt, D., Hinrichs, C., Wittenberg, T., Daumke, P., and Oppelt, P. G. 2016. The Clinical Data Intelligence Project. Informatik Spektrum 39, 4, 290–300.

[20] Alexander Bondarev. tMDataLoader. https://github.com/ThomsonReuters-LSPS/tMDataLoader. Accessed 15 March 2017.

[21] Turnbull, J. 2013. The Logstash Book. James Turnbull.

[22] Smith, J. J., Deane, N. G., Wu, F., Merchant, N. B., Zhang, B., Jiang, A., Lu, P., Johnson, J. C., Schmidt, C., Bailey, C. E., Eschrich, S., Kis, C., Levy, S., Washington, M. K., Heslin, M. J., Coffey, R. J., Yeatman, T. J., Shyr, Y., and Beauchamp, R. D. 2010. Experimentally derived metastasis gene expression profile predicts recurrence and death in patients with colon cancer. Gastroenterology 138, 3, 958–968.

[23] Hripcsak, G., Duke, J. D., Shah, N. H., Reich, C. G., Huser, V., Schuemie, M. J., Suchard, M. A., Park, R. W., Wong, I. C. K., Rijnbeek, P. R., and others. 2015. Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. Studies in health technology and informatics 216, 574.

[24] He, S., Yong, M., Matthews, P. M., and Guo, Y. 2016. tranSMART-XNAT connector-image selection based on clinical phenotypes and genetic profiles. Bioinformatics (Oxford, England).

[25] Wagholikar, K. B., Mandel, J. C., Klann, J. G., Wattanasin, N., Mendis, M., Chute, C. G., Mandl, K. D., and Murphy, S. N. 2016. SMART-on-FHIR implemented over i2b2. Journal of the American Medical Informatics Association : JAMIA.

[26] Sedlmayr, M., Wurfl, T., Maier, C., Haberle, L., Fasching, P., Prokosch, H.-U., and Christoph, J. 2016. Optimizing R with SparkR on a commodity cluster for biomedical research. Computer methods and programs in biomedicine 137, 321–328.