frontiers in Cell and Developmental Biology

PERSPECTIVE
published: 17 December 2019
doi: 10.3389/fcell.2019.00349

Ensuring Quality Standards and Reproducible Research for Data Analysis Services in Oncology: A Cooperative Service Model

Frank Emmert-Streib[1,2]*, *Matthias Dehmer*[3,4,5] *and Olli Yli-Harja*[2,6]

[1] *Predictive Society and Data Analytics Lab, Faculty of Information Technology and Communication Sciences, Tampere University, Tampere, Finland,* [2] *Institute of Biosciences and Medical Technology, Tampere, Finland,* [3] *Steyr School of Management, University of Applied Sciences Upper Austria, Steyr, Austria,* [4] *Department of Mechatronics and Biomedical Computer Science, UMIT, Hall in Tyrol, Austria,* [5] *College of Artificial Intelligence, Nankai University, Tianjin, China,* [6] *Institute for Systems Biology, Seattle, WA, United States*

Modern molecular high-throughput devices, e.g., next-generation sequencing, have transformed medical research. Resulting data sets are usually high-dimensional on a genomic-scale providing multi-factorial information from intertwined molecular and cellular activities of genes and their products. This genomics-revolution installed precision medicine offering breathtaking opportunities for patient's diagnosis and treatment. However, due to the speed of these developments the quality standards of the involved data analyses are lacking behind, as exemplified by the infamous Duke Saga. In this paper, we argue in favor of a two-stage cooperative serve model that couples data generation and data analysis in the most beneficial way from the perspective of a patient to ensure data analysis quality standards including reproducible research.

Keywords: computational biology, biostatistics, genomics, reproducible research, oncology, precision medicine, data science

OPEN ACCESS

Edited by:
Sol Efroni,
Bar-Ilan University, Israel

Reviewed by:
Geir Kjetil Sandve,
University of Oslo, Norway
Stephanie Roessler,
Heidelberg University, Germany

***Correspondence:**
Frank Emmert-Streib
v@bio-complexity.com

Specialty section:
This article was submitted to Molecular Medicine, a section of the journal Frontiers in Cell and Developmental Biology

Received: *21 September 2019*
Accepted: *04 December 2019*
Published: *17 December 2019*

Citation:
Emmert-Streib F, Dehmer M and Yli-Harja O (2019) Ensuring Quality Standards and Reproducible Research for Data Analysis Services in Oncology: A Cooperative Service Model. Front. Cell Dev. Biol. 7:349. doi: 10.3389/fcell.2019.00349

1. INTRODUCTION

Every new era provides opportunities but also challenges. For instance, at the early stage of the Industrial Revolution several severe accidents happened, one of which was the explosion of a steam boiler at a brewery in Mannheim in 1865. This and similar incidents resulted in the establishment of the TÜV (english meaning is Technical Inspection Association) in Germany as an independent institution for providing inspection and product certification services. Since then from a high-tech nuclear power plant to a simple hair-dryer every factory, product or service needs to pass a mandatory safety test from this independent association before it can be put into operation. Different countries may have different implementation rules and enforcing institutions but essentially every western country follows this model for all produces and services. However, there is one notable exception to the above and this relates to the analysis services of biomedical data.

An unfortunate example that demonstrates the catastrophic consequences of the lack of quality control standards in data analysis services in genomic medicine is the Duke Saga (Kolata, 2011). Specifically, a re-analysis of cancer genomic studies conducted at the Duke University by Anil Potti by external experts (Keith Baggerly and Kevin Coombes) from the MD Anderson found that various publications contained fundamental flaws and even scientific misconduct

(Baggerly and Coombes, 2009; Potti et al., 2011). These issues were so severe leading ultimately in the discontinuation of three clinical cancer trials that were started as a consequence of Anil Potti's research findings and the retraction of over ten scientific papers, all published in renowned journals, including the New England Journal of Medicine, Lancet Oncology and Nature Medicine. Sadly, further examples along these lines are ample (Ioannidis, 2005; Godlee et al., 2011; Simmons et al., 2011; Gupta, 2013; Tripathi et al., 2013).

2. MEASURES FOR ENSURING QUALITY STANDARDS

In our opinion the Duke Saga has similarities to the steam boiler explosion in Mannheim and we need to draw similar consequences from this incident. Specifically, we suggest that data analysis processes applied to medical, clinical and biomedical data, from which conclusions are drawn that will be used for the diagnosis or treatment of patients, need to be approved and certified by an external association in order to minimize the risk for patients. Here two important parameters of such an external association are:

(A) The independence of the external association from the data generating institution.
(B) The demonstrated expertise of the members constituting the external association.

2.1. Independence of the External Association From the Data Generating Institution

With respect to point (A), the independence of the external association from the data supplying institution needs to include its financial independence. This is in fact a problem with the current system. Specifically, nearly every medical or clinical institution maintains nowadays departments for biostatistics or computational biology. However, the scientists employed in these departments can hardly make decisions that are not aligned to the strategic interests of the institution. In contrast, an external association that is financially independent doesn't have to consider such strategic directions, in fact, it must not consider these because flawed decisions can endanger the well-being of patients.

2.2. Demonstrated Expertise of the Members Constituting the External Association

With respect to point (B), it is important that the members of the external association have a PI status. This ensures the highest possible standards of the quality control service that is needed in a clinical context. This is necessary because the fast paced developments in cancer genomics and its data analysis processes require constantly novel solutions that are not available off-the-shelf (Dunn and Bourne, 2017; Emmert-Streib and Dehmer, 2019).

This is also in contrast to most biostatistics and computational biology departments at medical or clinical institutions, which serve frequently merely as facilities to provide support for other departments at the same institution. This implies also that such facilities usually don't have a budget for developing new methods or for finding these, e.g., by a comparative analysis.

2.3. Similarities to the FDA

A related organization that has some similarities to our envisioned external association is the Food and Drug Administration (FDA) of the United States Department of Health and Human Services. For instance, the FDA is involved in the approval of medications, vaccinations and cosmetics. However, in contrast to our model described above, the FDA is predominantly concerned about outcome rather than the process leading to an outcome. That means the FDA does not perform experiments or computational analyses neither does the FDA provide such services. Our external association operates on a finer scale involved also in the process that leads to the outcome.

3. CONNECTION BETWEEN QUALITY STANDARDS AND REPRODUCIBLE RESEARCH

A problem that is tightly connected to ensuring quality standards of data analysis results is reproducible research (Jasny et al., 2011). In recent years, it has been recognized that in times of increasing usage of advanced technologies for the generation of data, requiring also more sophisticated data analysis methods, ensuring the reproducibility of such studies is far from being trivial. In fact, many studies have been identified that are lacking this important requirement (Ioannidis et al., 2009; Nekrutenko and Taylor, 2012). For this reason, minimal standards have been established that should be followed (Sandve et al., 2013) and important elements of such standards include:

- documentation of all steps (data generation and data analysis).
- archive analysis software (including version control).
- store seed of random number generator.
- provide access to the analysis pipeline.

As a simple test for the reproducibility of a study it is often informative to ask a colleague from a different department to repeat the analysis, as described in the documentation. This may reveal problems with different versions of software (e.g., R packages), gaps in the documentation or inconsistencies in the preprocessing of the data. Hence, even such a simple test can be very helpful in spotting problems.

It is obvious that problems with the reproducibility of studies could be easily avoided by the involvement of an external association because general quality standards of a data analysis include the requirement of its reproducibility. This underlines that a wider look to a problem can be very beneficial because it can lead to the resolution/avoidance of related problems.

4. PRACTICAL IMPLEMENTATION

So far we discussed the problem from a principle point of view. Now we turn to the practical implementation. In **Figure 1**, we summarize our discussion by providing a graphical visualization of the interplay between a data generating institution (outlined in blue) and an external association (outlined in orange). Here we distinguished visually between two components that are both part of the external association. The first involves all practical aspects of the data analysis including preprocessing and data integration, whereas the second one is only concerned with conclusions and recommendations derived from such an analysis.

Given the fact that many data generating institutions have already facilities for the analysis of data, e.g., biostatistics units, it would be efficient to utilize these in the following way. Instead of leaving it to the facilities to decide how to analyze the data, the external association should establish certified analysis protocols to follow. That means instead of performing the data analysis externally, it could be performed internally by the data generating institution itself, however, by following strict guidelines. In this way the lack of a research budget of facilities to establish optimal analysis protocols is compensated by the expertise of the external association.

As a side-effect, this would also deal with the problem of reproducibility because this is part of overall sound quality standards. Hence, only the part of the data analysis concerned with conclusions and recommendations should be under the sole governance of the external association.

5. HIGHER STANDARDS BY A COOPERATION SERVICE MODEL

We would like to emphasize that we consider the interplay between a data generating institution and an external association as a *cooperation*. The reason is that in research areas involving patients, the well-being and interests of the patients are top priority. This implies that it is not possible to keep analysis protocols shut way. In turn this means it will always be possible to reveal potential shortcomings or errors, as accomplished by Baggerly and Coombes (2009) and (Potti et al., 2011), because if undiscovered such errors will otherwise lead to physical or psychological harm of patients. Hence, in order to achieve the best possible outcome for the patients a cooperation between all involved parties is required that share responsibilities.

6. QUALITY STANDARDS AND REPRODUCIBLE RESEARCH BEYOND ONCOLOGY

The above discussion centered around cancer genomics data and the Duke Saga which happened in oncology (involving breast, colon, ovarian and lung cancer). However, we are of the opinion that such a collaborative model between a data generating institution and an external association as outlined above should be also beneficial for fields other than oncology.

A general characteristics of precision medicine and personalized medicine is that advanced data generation technologies are utilized in combination with sophisticated data analysis methods (Auffray et al., 2009; Ginsburg and Willard, 2009; Emmert-Streib and Dehmer, 2018). Since this is very similar to cancer genomics a corresponding translation of our outlined model should be transferable to other disease domains of precision medicine, e.g., immunology, neurodegenerative diseases or diabetes.

A current example that adds to our argument is given by Zolgensma. Zolgensma is an FDA approved gene therapy by Novartis intended to treat children with spinal muscular atrophy (SMA). This is the most severe form of SMA. On June 28 2019 the FDA was informed by AveXis about a data manipulation issue during product testing (FDA, 2019; Tirrell, 2019). This is also an example demonstrating the severity of the problem beyond oncology raised in this paper that can effect the life of patients and even children. Another example is the ban of the European Union of around 700 generic medicines for alleged manipulation of clinical trials conducted by the company GVK Biosciences (EMA, 2015).

7. DISCUSSION

Due to the difficulty of the raised problem, we would like to clarify some further issues. First, our arguments are limited to biomedical studies involving either directly or indirectly patients. However, our arguments do not extend to general biological studies. The reason for this is that here we are not concerned with general reproducibility problems of studies but with consequences of such issues for patient treatment and care. That means, in our opinion, we focus on the most severe problem we are currently facing. If our arguments should also be extended to biological studies is open for discussion.

The crucial point is whether the outcome of an analysis from either diagnostic, prognostic or predictive investigations is affecting the treatment of patients with a clear causal connection between both. That means the analysis needs to inform the treatment. Hence, for instance an analysis of patient-derived cell lines, which only indirectly involves patients, falls under this category if the outcome of this analysis has a direct influence on the patient's treatment. It is clear that much of biomedical research is not directly concluding on any issue related to patient treatment, but instead typically investigates biological mechanisms.

Second, it is clear that the translation of biological findings toward their clinical usability is a long and difficult endeavor. As an example, we would like to mention the known difficulties of finding reproducible prognostic methylation biomarkers for colorectal cancer (Draht et al., 2018) or general cancers (Koch et al., 2018). This just underlines the difficulties we are facing experimentally and computationally requiring stringent protocols to safeguard against spurious results.

Third, in order to make an impact, we belief it is necessary to specify our scope precisely. The problem is that precision medicine or personalized medicine could refer to highly variable

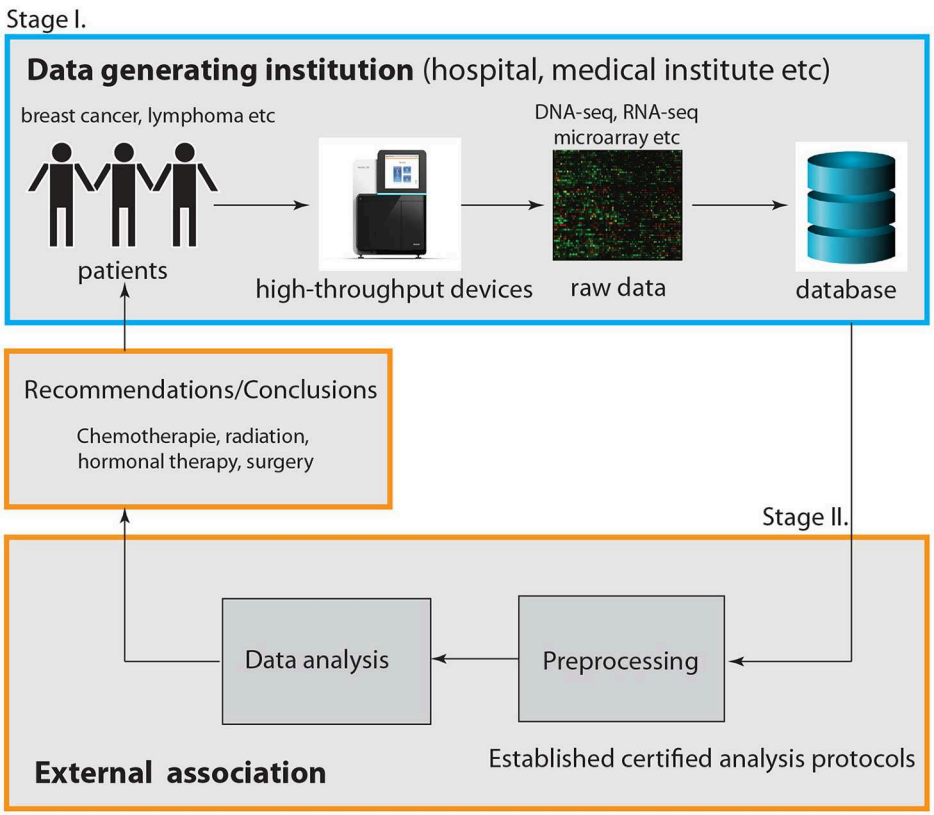

FIGURE 1 | Visualization of the interplay between a data generating institution, e.g., a hospital or a medical institution, and an external association. Overall, this represents a two-stage cooperative service model.

settings, ranging from basic research employing patient samples to decipher disease biology to drug development or clinical assessment. Unfortunately, these examples show that defining a scope precisely is not straight forward since all of these sub-studies relate to *"patients"*. For this reason, we suggest an assessment of the outcome of a study if it can potentially *"harm a patient."* In this way the problem is converted to a legal issue and its definition is given by country-specific laws.

Fourth, data privacy is a current issue of great relevance in the big biomedical data era (Malin et al., 2013; Patil and Seshadri, 2014). However, our focus is on preventing patient harm due to inadequate data analysis standards. Of course, patients could also suffer from data privacy violations by third parties, however, not due to inadequate data analysis standards. Hence, data privacy is an issue data analysis services need to adhere to but for different reasons.

Fifth, erroneous data analysis results could come from deliberate cheating or other forms of scientific flaws. Interestingly, from a patient perspective, the potential harm is the same. However, safeguarding against the former threat is naturally accomplished by an external association because many of the incentives are eliminated in this way.

Sixth, problems of the sort discussed in this paper are actually rather widespread. For instance, in a survey study conducted in Bozzo et al. (2017) the authors identified 571 retracted publications in the cancer research literature of which 28.4% of the retrations were due to fraud and 24.2% due to errors. Further examples are provided in George and Buyse (2015) where the fabrication or falsification of data in clinical trials has been investigated effecting hundreds of publications in the literature.

8. CONCLUSIONS

A necessary step toward the practical realization for the certification of medical and clinical data analysis services would require the authorities to become active. For instance, legal laws could be legislated making the acquisition of such certificates mandatory attesting the fulfillment of quality standards. In turn, this would enable the establishment of external associations, similar to the model of the TÜV. If the Duke Saga could have been prevented is speculation. However, given the fact that Baggerly and Coombes (2009) were capable of reverse engineering some errors it appears reasonable to assume that an external association could have picked up these at an early stage before entering clinical trials.

In MacArthur (2012) it is argued that "Flawed papers cause harm beyond their authors: they trigger futile projects, stalling the careers of graduate students and postdocs, and they degrade

the reputation of genomic research." Importantly, we would like to add that flawed medical data analysis services can even severely harm the life of patients. In summary, we argued in favor of a "trust, but verify" approach because since the Duke Saga no legislative changes were initiated allowing history to repeat.

AUTHOR CONTRIBUTIONS

All authors contributed to all aspects of the preparation and the writing of the manuscript.

FUNDING

MD thanks the Austrian Science Funds for supporting this work (project P30031).

ACKNOWLEDGMENTS

FE-S would like to thank the participants of the 62. Annual Conference of the German Society for Medical Informatics, Biometrics and Epidemiology (GMDS) (Oldenburg, Germany) and especially Rainer Röhrig and Harald Binder for fruitful discussions.

REFERENCES

Auffray, C., Chen, Z., and Hood, L. (2009). Systems medicine: the future of medical genomics and healthcare. *Genome Med.* 1:2. doi: 10.1186/gm2

Baggerly, K. A., and Coombes, K. R. (2009). Deriving chemosensitivity from cell lines: forensic bioinformatics and reproducible research in high-throughput biology. *Ann. Appl. Stat.* 3, 1309–1334. doi: 10.1214/09-AOAS291

Bozzo, A., Bali, K., Evaniew, N., and Ghert, M. (2017). Retractions in cancer research: a systematic survey. *Res. Integr. Peer Rev.* 2:5. doi: 10.1186/s41073-017-0031-1

Draht, M. X., Goudkade, D., Koch, A., Grabsch, H. I., Weijenberg, M. P., van Engeland, M., et al. (2018). Prognostic dna methylation markers for sporadic colorectal cancer: a systematic review. *Clin. Epigenet.* 10:35. doi: 10.1186/s13148-018-0461-8

Dunn, M. C., and Bourne, P. E. (2017). Building the biomedical data science workforce. *PLoS Biol.* 15:e2003082. doi: 10.1371/journal.pbio.2003082

EMA (2015). *GVK Biosciences: European Medicines Agency Recommends Suspending Medicines Over Flawed Studies*. European Medicines Agency. Available online at: https://www.ema.europa.eu/en/documents/referral/gvk-biosciences-article-31-referral-gvk-biosciences-european-medicines-agency-recommends-suspending_en.pdf

Emmert-Streib, F., and Dehmer (2018). A machine learning perspective on personalized medicine: an automatized, comprehensive knowledge base with ontology for pattern recognition. *Mach. Learn. Knowl. Extr.* 1, 149–156. doi: 10.3390/make1010009

Emmert-Streib, F., and Dehmer, M. (2019). Defining data science by a data-driven quantification of the community. *Mach. Learn. Knowl. Extract.* 1, 235–251. doi: 10.3390/make1010015

FDA (2019). *Statement on Data Accuracy Issues With Recently Approved Gene Therapy*. US Food and Drug Administration. Available online at: https://www.fda.gov/news-events/press-announcements/statement-data-accuracy-issues-recently-approved-gene-therapy?utm_campaign=080619_Statement_Statement%20by%20FDA%20on%20data%20accuracy%20issues%20with%20gene%20therapy&utm_medium=email&utm_source=Eloqua

George, S. L., and Buyse, M. (2015). Data fraud in clinical trials. *Clin. Investigat.* 5:161. doi: 10.4155/cli.14.116

Ginsburg, G. S., and Willard, H. F. (2009). Genomic and personalized medicine: foundations and applications. *Translat. Res.* 154, 277–287. doi: 10.1016/j.trsl.2009.09.005

Godlee, F., Smith, J., and Marcovitch, H. (2011). Wakefield's article linking MMR vaccine and autism was fraudulent. *BMJ* 342:c7452. doi: 10.1136/bmj.c7452

Gupta, A. (2013). Fraud and misconduct in clinical research: a concern. *Perspect. Clin. Res.* 4:144. doi: 10.4103/2229-3485.111800

Ioannidis, J. P., Allison, D. B., Ball, C. A., Coulibaly, I., Cui, X., Culhane, A. C., et al. (2009). Repeatability of published microarray gene expression analyses. *Nat. Genet.* 41:149. doi: 10.1038/ng.295

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Med.* 2:e124. doi: 10.1371/journal.pmed.0020124

Jasny, B., Chin, G., Chong, L., and Vignieri, S. (2011). Data replication & reproducibility. again, and again, and again.... introduction. *Science* 334, 1225–1225. doi: 10.1126/science.334.6060.1225

Koch, A., Joosten, S. C., Feng, Z., de Ruijter, T. C., Draht, M. X., Melotte, V., et al. (2018). Analysis of DNA methylation in cancer: location revisited. *Nat. Rev. Clin. Oncol.* 15:459. doi: 10.1038/s41571-018-0004-4

Kolata, G. (2011). *How Bright Promise in Cancer Testing Fell Apart*. The New York Times. Available online at: http://www.nytimes.com/2011/07/08/health/research/08genes.html?_r=0

MacArthur, D. (2012). Methods: face up to false positives. *Nature* 487, 427–428. doi: 10.1038/487427a

Malin, B. A., El Emam, K., and O'Keefe, C. M. (2013). Biomedical data privacy: problems, perspectives, and recent advances. *J. Am. Med. Informat. Assoc.* 20:2. doi: 10.1136/amiajnl-2012-001509

Nekrutenko, A., and Taylor, J. (2012). Next-generation sequencing data interpretation: enhancing reproducibility and accessibility. *Nat. Rev. Genet.* 13:667. doi: 10.1038/nrg3305

Patil, H. K., and Seshadri, R. (2014). "Big data security and privacy issues in healthcare," in *2014 IEEE International Congress on Big Data* (Anchorage, AK: IEEE), 762–765.

Potti, A., Dressman, H. K., Bild, A., Riedel, R. F., Chan, G., Sayer, R., et al. (2011). Retraction: genomic signatures to guide the use of chemotherapeutics. *Nat. Med.* 17:135. doi: 10.1038/nm0111-135

Sandve, G. K., Nekrutenko, A., Taylor, J., and Hovig, E. (2013). Ten simple rules for reproducible computational research. *PLoS Comput. Biol.* 9:e1003285. doi: 10.1371/journal.pcbi.1003285

Simmons, J. P., Nelson, L. D., and Simonsohn, U. (2011). False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol. Sci.* 22, 1359–1366. doi: 10.1177/0956797611417632

Tirrell, M. (2019). *Novartis Fires Brother Scientists Alleged to be Involved in Data Manipulation*. CNBC. Available online at: https://www.cnbc.com/2019/08/14/novartis-allegedly-fires-brother-scientists-over-data-manipulation.html

Tripathi, S., Glazko, G., and Emmert-Streib, F. (2013). Ensuring the statistical soundness of competitive gene set approaches: gene filtering and genome-scale coverage are essential. *Nucl. Acids Res.* 6:e53354. doi: 10.1093/nar/gkt054

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.