Towards a More Data Oriented Medical Research Environment

Survey Results on Information and Data Practice

Lars Müller, Christoph Szepanski, Thomas Wetzel, Hans-Christoph Hobohm[1]

University of Applied Sciences, Faculty of Information Sciences
Friedrich-Ebert-Str. 4, 14467 Potsdam, Germany
hobohm@fh-potsdam.de

Abstract

Objective: The article contains results from a survey conducted among medical professionals regarding the way in which they handle research data in practice. The aim of the survey was to identify potential for improved data orientation and to ascertain starting points for support tools. The results were to be used to develop creativity-promoting computer-based tools.
Method: The empirical basis was provided by guideline-based qualitative interviews and a quantitatively structured online survey. The interviews were partly transcribed and evaluated using content analysis.
Results: Data analyses are primarily conducted in a target-oriented manner, i.e. on the basis of a hypothesis, with data centres representing an important source of reference for the surveyed medical professionals. The means used for analysing data indicate individual working practices, with the analyses often being conducted at the medical professionals' desks and towards the end of the working day. The results of analyses are often used in publications but less used for research applications.

---

1 corresponding author

In: F. Pehar/C. Schlögl/C. Wolff (Eds.). Re:inventing Information Science in the Networked Society. Proceedings of the 14th International Symposium on Information Science (ISI 2015), Zadar, Croatia, 19th–21st May 2015. Glückstadt: Verlag Werner Hülsbusch, pp. 230–243.

Conclusion: The results lead to the conclusion that methods for exploring data that are as easily accessible as possible with as few barriers as possible need to be offered. A prerequisite for a shift from hypothesis-oriented research to data-oriented research would appear to be a link with data centres that offer processed/prepared and anonymised data. Further data sources and other secondary information sources should also be able to be integrated. Complex visualisations can likewise support data-oriented working practices if they are offered as an option.

Keywords: Research data usage, Information behavior, Medical science, Information practice, Scientific methodology

1 Introduction

In the R&D project 'Data Creativity Tools for Innovation and Research' (DCT) at University of Applied Sciences Potsdam a web application has been created that is designed to foster the development of new research questions and hypotheses from large heterogeneous databases of medical research data. In cooperation with the project partner 'Open European Nephrology Science Center' (OpEN.SC) of Charité Berlin, an information environment has been designed to support the data analysis and identification of relevant knowledge gaps. Qualitative interviews and an online survey on the handling of medical research data in practice were conducted in order to make the tools as user-oriented as possible.

The aim of the survey was to find out more about how medical professionals deal with "their" data in practice. The online questionnaire was based on the following question:

At which points in the problem-finding and problem-naming process (data exploration, anomaly recognition, hypothesis formulation) can modified forms of data presentation, data integration and data networking help researching pathologists/nephrologists to make better use of their creative potential when formulating new research questions and hypotheses?

2 Literature review

Data orientation is nothing new in biomedical research observe Kell and Oliver (2004: 101). Nevertheless, they note that clear priority is accorded to hypothesis-oriented research. They argue in favour of a complementary balance between hypothesis-oriented research and data-oriented research in which the hypothesis is the result and not the starting point. "The mere generation and dissemination of data (…) is then seen – when viewed in the correct context – as a highly valuable component of the scientific process, even when no hypothesis was involved in the generation of those data" (Kell & Oliver 2004: 103). Bell et al. (2009) regard data-intensive science as an import element of future research. According to them, the development of special tools is necessary that enable and support it. Thessen and Patterson (2011) share this view in their detailed review of "Data issues in the life sciences". Furthermore, they conclude that the successful transformation towards data-oriented biology requires technical progress and a change in culture in the sub-disciplines.

This raises the question as to how this change in culture can be technically supported in order to generate hypotheses from data. Various authors have indicated from the perspective of creativity research that the problem finding process is an important step in generating innovations and that it has so far been unable to be automated (Nickerson 1999: 395; Runco 2014: 18; Cropley & Cropley 2009: 252 ff.). The science philosopher Mario Bunge adopts a similar position, identifying problem finding as a core part of scientific activity (Bunge 1998: 187) and observing "Yet man alone invents new problems: he is the only problemizing being, the only one which can feel a need to and a pleasure in adding difficulties to those posed by the natural and social environment" (Bunge 1998: 188). Hoover and Feldhusen have furthermore identified a connection between problem finding processes and information processes (Hoover & Feldhusen 1994: 213 f.).

Müller et al. (2012b; 2012a) have developed an ideal-type process model of the data-oriented problem finding process. It integrates approaches to developing data-oriented research and the creative generation of innovative hypotheses.

3 Methods

3.1 Operationalisation/design

The survey guidelines for the interviews were designed to ascertain information and communication behaviour, cooperation, research data, creativity and problem-solving skills. The operationalisation for the online survey was based on the premiss of a data-oriented research process (cf. Müller et al. 2012a) which begins with data analysis and ends with a new research project. Any differences and similarities between this model and practice would indicate possible starting points and areas for intervention in the research process.

A total of seven topic areas were surveyed:

1. Reason for data analysis to identify the extent of explorative data analysis practices
2. Choice of data sources
3. Application of data analyses to establish/ascertain data orientation
4. Use of external aids to identify the knowledge environment during data analysis
5. Preferred form of data representation
6. Work situation to indicate functional requirements
7. Professional experience background of the surveyed individuals to evaluate their answers.

Questions were asked as to how data are handled in practice. Results from these explorative interviews, broadly focussing on information activities in everyday medical practice, were used for the development of the subsequent survey which addressed the data handling and information gathering in more detail.

3.2 Data collection

The guideline-based interviews were conducted between 13.7.2011 and 25.8.2011. A total of five medical professionals were surveyed in individual interviews lasting approximately thirty minutes. The five participants (two were male and one female) are all researchers and/or practicing medical professionals who have completed their training and work in various specialist fields in medicine.

The subsequent online questionnaires were created and conducted via the online service http://www.unipark.info. The surveyed individuals accessed the questionnaire by clicking on a link. They were contacted using an internal e-mail distribution list of the nephrology department of Charité Berlin and were asked to participate in the survey (mixed self-recruitment and quota sample due to the degree of specialisation required). The answers were recorded in the period from mid-June to mid-September 2012. In order to evaluate their answers, we also asked the online survey participants about their professional background. Only two of the eight participants were new to their professional field. Six of those surveyed stated that they have a qualification as a specialist physician/consultant. The listed specialist fields were nephrology and internal medicine, with non-medical practitioners listing medical informatics and pharmaceutics. The numbers of cases were too low to observe differences in handling data according to years of practice or degree of qualification. Based on the information regarding qualifications, we can conclude that the survey participants had extensive experience in dealing with medical data and in the majority of cases many years of experience. When evaluating the answers, in some cases it appeared to be necessary to distinguish between the groups of clinical physicians vs. non-physicians, since questions that related to the direct treatment of patients are not relevant for non-physicians due to the nature of their particular professional role.

4 Results

4.1 Reasons for data analysis

The main reason why the researchers conduct this study is because they wish to pursue an idea they have had in mind. Research questions, the slightly more elaborate version of ideas, also provide a reason for data analysis (cf. fig. 1). Inspiration from scientific literature plays a lesser role. The exploration of data on the basis of unfocussed interest is the exception. Wanting to pursue ideas is the reason for data analysis rather than the motivation to find ideas. The answers show that data analysis tends to be hypothesis-based rather than explorative. The results that certainly cannot be generalized due

to the rather small sample might be specific to the medical domain under observation here.

4.2 Selection of data sources

When asked about the choice of data source, it is noticeable that re-using data from colleagues is never listed as the most frequent option. Surprisingly, half of the survey participants listed data centres as a frequently used source, which in fact is a form of re-use. The answers relating to data centres probably refer to the various data centres available within the Charité and which represent sub-elements of the hospital information system. This includes, for example, parts of the hospital's internal SAP system, the Charité intranet or the parameter and reference value database provided by the Institute of Laboratory Medicine. Here, the different data sources do not permit a clear delineation from the re-use of colleagues' data. However, we suspect that the advantage of data centres can be seen in the fact that data are generally prepared and documented in a standardised way and that (data protection) legal issues regarding access and usability are governed by a specified procedure. Targeted collection of new data was cited as being more frequently used than existing patient data.

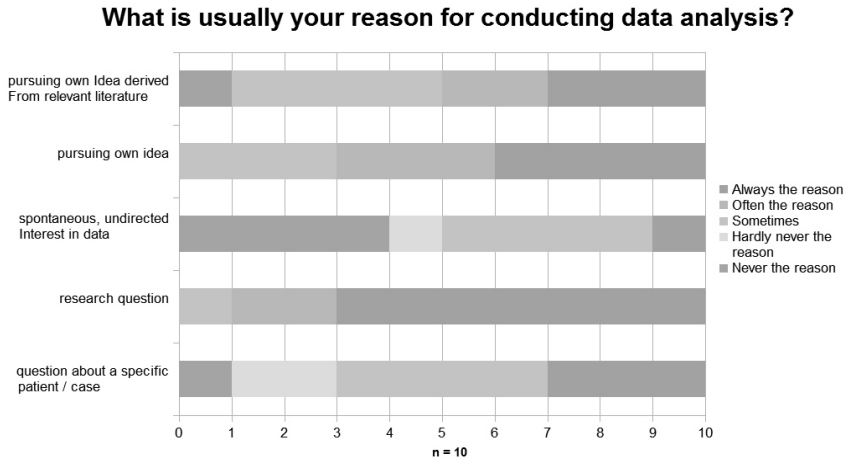

Figure 1. Results of the survey indicating the reasons for data analysis.

The qualitative interviews showed that there is particularly a demand for simple, safe and comprehensible access to research data. This was once again confirmed in the online survey, indicating that potential especially exists in the preparation of patient data in a data centre, as offered by OpEN.SC. In qualitative terms, the potential particularly lies in improved integration of data from different sources (other locations, other sub-disciplines).

4.3 Use of external resources

All forms of external information sources are important, scientific information media are clearly dominant. PubMed and journals are most frequently listed, however an important role is also played by unspecified Internet usage ("Google etc."). Reference tools for scientific publications enable the academic environment to be explored for certain questions. Google-Scholar also provides links to documents, but is used with a greater level of reservation. Only one surveyed person said that they frequently used this service. The intensity with which scientific reference systems are used is broadly scattered. The purpose for which general Internet usage was listed by all of the survey participants as "sometimes" or "frequently" was not included in the survey. However, answers from the guided interviews indicate that new developments are being searched for. The goal of supportive information systems should thus be to integrate as many important secondary information sources as possible. To expand the knowledge basis on certain ideas it is also helpful to integrate and actively offer less used tools. This produces a range of a) content depth, b) currency and c) general context. Interpersonal communications occupies a mid-table position and therefore appears to be an unexploited resource for data analysis.

4.4 Forms of displaying data

Traditional visualisations in the form of diagrams and curves are clearly preferred by most of survey participants. All survey participants prefer a table with numerical values at least partly. However, attitudes towards complex visualisations are much divided: four survey participants answered "not at all" or "hardly at all". In three cases, in contrast, this is one of the preferred forms ("yes definitely"). Complex visualisations increase the complexity of the displayed data. For complex questions this is no doubt necessary and in-

tended. At the time however, the interviews showed that there is little or no time and willingness for learning how to use new tools (new software). Nevertheless, this is precisely where the potential lies for developing creativity-promoting tools that encourage data-triggered hypothesis generation. Numerical data in tables or two- to three-dimensional graphics are not always able to adequately represent the complexity of data and their relationships. In traditional graphics and table forms, it is difficult if not impossible to visualise non-numerical values. Potential exists in new visualisation techniques whose value has perhaps not yet been recognised. As they were only listed as the preferred form of data presentation by three survey participants, they must be used with caution, i.e. optionally and ideally integrated in a way that will increase acceptance.

4.5 Work situation

Data analyses are conducted everywhere, being it at work, at home or during travelling – and at all times. There are, however, clear trends: physically at work and time-wise after working hours during leisure time – in other words probably outside of workplace routines. All practicing medical professionals stated in the survey that they also analyse data in their spare time, unlike those surveyed from the medical informatics/pharmaceutics professions. PCs are an essential tool, paper forms are still used. Tools that are more suited for cooperative working techniques such as projectors or whiteboards were not listed by anyone as a used tool. The answers indicate a location-independent, in other words web-based development, without high access barriers and which also includes mobile use. The question of access rights must therefore be clarified with the lowest possible technical access barriers for authorised users.

4.6 Use of analysis results

When asked how the results of data analyses are first used, formal publications are clearly top listed, followed by preparation for presentations. Conversations with colleagues about data analyses are important, but do not appear to be of central importance. Based on the qualitative interviews, it can be added that personal conversations with colleagues (face to face or by telephone) are the first choice when asked for specific instances. They take place

among things in the context of further training at conferences or work groups and in practice-internal or department-internal communication. Answers with regard to research applications, clinical findings and patents are broadly scattered. Data analyses are clearly carried out for the primary purpose of research – in other words to find or justify answers. Problem-finding or hypothesis-finding from data analyses could be increased by encouraging direct communication that encompasses the unfinished search.

5 Summary

The data analyses carried out by the survey participants are primarily conducted on a target-oriented basis, in other words based on a hypothesis. Data analysis is an individual process in which traditional working methods are favoured. Rather than being explorative in nature, the analyses are performed on a result-oriented basis with a view to being able to make substantiated statements, e.g. in publications. The potential which exists in the high level of data affinity among those surveyed is partly limited by technical and administrative barriers.

- The reason for data analysis is usually an (own) idea, while searching for an idea is hardly ever the trigger.
- Data centres are an important data source for those surveyed.
- Data analyses are conducted incorporating a broad knowledge pool: published scientific/academic literature and Internet research on current developments are used.
- Attitudes towards complex visualisations are mixed, with some people rejecting them and others preferring them.
- The tools used indicate that individual working methods are used for data analysis.
- Data analyses usually take place in the workplace and towards the end of the working day.
- The results of data analyses are primarily used in publications and less so for research applications.

5.1 Discussion

In addition to the direct benefits for DCT development, the survey results are of interest to us from a general information science point of view. A number of these aspects are discussed in the following. As to the question of why PubMed plays a prominent role as an information resource for our survey participants while other tools, such as Google Scholar, have only a secondary status as research tools for collecting evidence for scientific publications and why personal discussions with colleagues in the field are only ranked with medium importance, it may be useful to consider Zipf's Principle of Least Effort (PLE) (cf. Case 2005: 291). According to this theory, individuals will always seek to minimise the workload involved when searching for information, even though they know that the research results will be of a lower quality. The survey results and the interviews show that PubMed serves as the primary research resource. The clear preference can result in a problem for the user, which is also explained by PLE. If the user believes that he or she has found sufficient research results in PubMed, there is no incentive to conduct structured literature research that includes other resources, thus improving the quality of the research results. In the survey we only asked about people's preferences with regard to research tools, but not about the way in which they are used. Further, the clear favouring of a single tool (PubMed) could indicate that an insufficient knowledge base is being used, although the researchers have the subjective impression of finding everything they need to interpret the data (cf. Lawrence 2008: 83). Should this assumption be true, this would represent a major problem for research. It would also pose a challenge for information systems seeking to at least encourage broader coverage of potential sources in research. Zipf's PLE can also go some way towards explaining why discussions with colleagues are only occasionally used an external data source. If we take into account the most frequently listed answers as to the time (towards the end of the working day) and place (in the workplace) when data analysis is conducted, this could allow us to conclude that here, too, individual opportunity costs are weighed up, e.g. the understandable consideration during everyday work in medicine of whether the research context should be extensively explained to the colleague beforehand – added to the fact that the colleague must be accessible at the time when the inspiration strikes – or whether the time needed for this would be better invested in carrying out research work. A conventional web-based information system can exercise little influence on the time and environment in which

data analysis is performed. On the question of improved integration of information systems in daily work, Raya Fidel recommends the Cognitive Work Analysis approach (2012: 225 f.). Pursuing this approach could also open up new avenues for application development in terms of how the exchange of knowledge between colleagues can be better incorporated into data-oriented research processes. The re-use of data plays a less important role for our survey participants. This observation may well be characteristic for medical research as a whole. An important basis for the re-use of data is formed by easy accessibility, reliability of both the users and systems and recognition of the work carried out by those who provide the data for re-use (cf. Tenopir et al. 2011: 9 ff.).

5.2 Limitations

With precisely 10 fully completed questionnaires the return rate fell far short of expectations. As we can assume with a high level of probability that the answers come from the core of target group due to the automatically recorded web statistics, conclusions can still be drawn despite the low return rate. However, there is no question that the survey results are of an explorative nature. Reliable statements regarding the data practices of medical professionals can only be made after repeating the survey and achieving a significantly higher return rate. Nevertheless the findings confirm generally the research in information behaviour in other fields, e.g. concerning the information source of the first choice being the nearby colleague (cf. Hobohm 2013). The negligence of the explorative aspect of information handling on the other hand is slightly against current information science discourse (Dörk et al. 2011) and asks for further research too. The rather small amount of literature on the specific subject of data and information practice in this medical field indicates a certain necessity for further research.

Another limitation of the results is that the participants were not directly observed handling data in their daily work and the survey relied on information provided by the participants. It is therefore likely that implicit processes or reciprocal effects when handling data remain undetected. In the light of the practice turn (Huizing & Cavanagh 2011) of the information science other investigation methods should be used.

6 Conclusion

Our premiss says that for data-oriented research the idea must come from the "spontaneous, undirected interest" in the data, in other words entirely from the purpose of exploration.

Through the development of creativity-promoting tools (Müller et al. 2012a), a "short circuit" needs to be triggered here in which data exploration generates ideas for formulating new hypothesis. These ideas create reasons for the data to be analysed again in a new light – and in the subsequent research process the data will then be analysed on a hypothesis-led basis.

To this end, as many low-barrier methods as possible for data exploration (both structure and data values) need to be offered. This will enable ideas to be created that will lead to a more intensive examination of the available data. Close ties with data centres providing processed and anonymised data is beneficial, as here the willingness for re-use is the highest. Systems for supporting data-oriented research should provide the opportunity to integrate further data centres if required. Applications in this area should also integrate a selection of secondary information sources that is as broad as possible and encourage their use in order to optimise the knowledge pool.

Innovative visualisations can help to deal with the increasing complexity of data. However, as strong reservations exist toward these visualisations, they should be offered as an option and not as the main or only tool for data exploration in the information system.

Contrary to other information seeking situations, discussions with colleagues play a less important role in the data analysis process. One reason for this could be that it is usually rather difficult to verbally communicate raw data and their context. We believe that discussions only take place once the data has been prepared, for example in the form of a diagram. For the development of suitable applications, this assumption means that there is no point in pursuing elaborate measures for supporting communication, for example implementing Social Media. However, the possibilities for strengthening the use of knowledge resources through personal exchange should also be taken into account in the form of analogue channels. For example, simple or complex visualisations and appropriate computer screen representations displayed via a projector could simplify a cooperative working approach.

As we have shown, the results from our survey allowed us to derive a number of practice-oriented ideas for supporting data-oriented research. Ex-

perimenting with developments in this area and testing their effects in practice appears to be a promising avenue to pursue. A prototype application of this nature has been developed in the context of the research project as the so called *Data-Creativity-Tools for Innovation and Research* (http://datacreativity.fh-potsdam.de/).

Acknowledgements

This research has been made possible by a research grant from the German Federal Ministry of Education and Research (BMBF). Grant number: 17089X10. – Translation provided by Natalie Chandler.

References

Bell, G, T. Hey, & A. Szalay (2009). Computer Science: Beyond the Data Deluge. *Science* 323, 1297–1298.

Bunge, Mario (1998). *Philosophy of science*. Rev. ed. New Brunswick, NJ: Transaction Publ.

Case, Donald O. (2005). Principle of Least Effort. In: Fisher, Karen E.; Erdelez, Sandra; McKechnie, Lynne (Eds.). *Theories of Information Behavior*. Cambridge, Mass. : MIT Press, pp. 289–292.

Cooke, Colin R, & Theodore J. Iwashyna (2013). Using Existing Data to Address Important Clinical Questions in Critical Care. *Critical Care Medicine* 41, 886–896.

Cropley, Arthur, & David Cropley (2009). *Fostering creativity. A diagnostic approach for higher education and organizations*. Cresskill, NJ: Hampton Press.

Dörk, Marian; Carpendale, Sheelagh; Williamson, Carey (2011). The Information Flaneur. A Fresh Look at Information Seeking. In: *CHI2011. Vancouver, May 7–12*. ACM.

Fidel, Raya (2012). *Human Information Interaction: An Ecological Approach to Information Behavior*. Cambridge et al.: MIT Press.

Hobohm, Hans-Christoph (2013). Informationsverhalten (Mensch und Information). In: Rainer Kuhlen, Wolfgang Semar & Dietmar Strauch (Eds.). *Grundlagen der praktischen Information und Dokumentation.* 6th Ed. Berlin: De Gruyter Saur, pp. 109–125.

Hoover, Steven M, & John F. Feldhusen (1994). Scientific Problem Solving and Problem Finding: A Theoretical Model. In: *Problem finding, problem solving, and creativity.* Creativity research, ed. by Mark A. Runco. Norwood N.J: Ablex Pub. Corp., pp. 201–219.

Huizing, Ard, & Mary Cavanagh (2011). Planting contemporary practice theory in the garden of information science. *Information Research* 16 (4).

Kell, Douglas B, & Stephen G. Oliver (2004). Here is the evidence, now what is the hypothesis? The complementary roles of inductive and hypothesis-driven science in the post-genomic era. *BioEssays* 26, 99–105.

Lawrence, David W. (2008). The information-seeking behaviors of professionals and information sources in the field of injury prevention and safety promotion. Stockholm: Division of International Health, Department of Public Health Sciences, Karolinska institutet.

Müller, Lars, Thomas Wetzel, Hans-Christoph Hobohm, & Thomas Schrader (2012a). Creativity Support Tools for Data Triggered Hypothesis Generation. In *2012 Seventh International Conference on Knowledge, Information and Creativity Support Systems (KICSS 2012)*, ed. by C. S. Lee & K. L. Ong. New York, NY: IEEE, pp. 24–27.

Müller, Lars, Thomas Wetzel, & Hans-Christoph Hobohm (2012b). Auf kreativen Wegen von Daten zum Wissen am Beispiel medizinischer Forschungsdaten. In *Vernetztes Wissen – Daten, Menschen, Systeme – WissKom 2012*, ed. by Bernhard Mittermaier, Jülich: Forschungszentrum Jülich, pp. 93–105.

Nickerson, Raymond S. (1999). Enhancing Creativity. In *Handbook of creativity*, ed. by Robert J. Sternberg. Cambridge: Cambridge Univ. Press, pp. 392–430.

Runco, Mark A. (2014). *Creativity. Theories and themes; research, development, and practice,* Burlington: Elsevier.

Tenopir, Carol; Allard, Suzie; Douglass, Kimberly; Aydinoglu, Arsev U.; Wu, Lei; Read, Eleanor; Maribeth Manoff, Mike Frame & Cameron Neylon (2011). Data Sharing by Scientists: Practices and Perceptions. In *PLoS ONE* 6, e21101.

Thessen, Anne, & David Patterson (2011). Data issues in the life sciences. *ZooKeys* 150, 15.