Evaluating the Usefulness of Mobile Services based on Captured Usage Data from Longitudinal Field Trials

Jensen, Kasper Løvborg; Larsen, Lars Bo

Published in:
Ikke angivet

Publication date:
2007

Document Version
Publisher's PDF, also known as Version of record

Evaluating the Usefulness of Mobile Services based on Captured Usage Data from Longitudinal Field Trials

Kasper Løvborg Jensen and Lars Bo Larsen
Multimedia Information and Signal Processing
Department of Electronic Systems
Aalborg University, Denmark
{klj,lbl}@es.aau.dk

ABSTRACT

It is widely discussed whether the usefulness of mobile systems and services should be evaluated in the field or in the lab. The aim of this paper is to make a case for conducting longitudinal field trials when evaluating the usefulness of mobile services, and to partly base such evaluations on quantitative usage data automatically captured on the device. A framework for automatic capture and analysis of such data is introduced, and it is shown how it has been used to evaluate a mobile service for diabetes management. Results and experiences based on a three month field trial with a single diabetic user is presented and it is demonstrated how these results can be used to make inferences about the use of the service with regard to functionality and usage patterns.

Categories and Subject Descriptors

D.2.2 [**Software Engineering**]: Design Tools and Techniques; H.5.2 [**Information Systems**]: User Interfaces—*Evaluation/methodology, Graphical user interfaces (GUI), Prototyping, Theory and methods*; J.3 [**Computer Applications**]: Life and Medical Sciences—*Medical information systems*

General Terms

Design, Experimentation, Measurement, Human Factors

Keywords

Usefulness, Evaluation, Mobile, Longitudinal, Field Trial, Usage Data, Automatic Data Capture and Analysis

1. INTRODUCTION

Most researchers agree on the importance of involving users when evaluating the usefulness of novel mobile systems and services. However, it is widely discussed whether such evaluations should be performed in the field or in the laboratory. Jakob Nielsen [14] defined usefulness as the combination of utility and usability, i.e. a service has to both be usable and have utility for the user in order to be useful. This distinction fits well with the current focus on User Experience design and represents a more holistic view on usability compared to previous trends. However, this distinction is somewhat abstract and it can sometimes be hard to attribute a user's attitude to one or the other when evaluating it. Currently, most user tests are conducted in laboratory setting with emphasis on usability, although they also to some degree include utility.

1.1 Lab vs. Field Evaluation

The primary arguments in favour of lab-based evaluation are the high degree of control and the relatively easy collection of data. Furthermore, there are many practical and logistic problems associated with field testing, which are much easier handled in the lab. In other words, field experiments are both more difficult and more costly with regard to time and manpower. This is why only a relatively small fraction of research within mobile HCI is based on field experiments [12].

The advantages of field experiments are that they will incorporate the natural context and environment of the user of the given service. When evaluating mobile services this can be extremely important. User interaction with mobile systems and services is very complex as it depends on the environment and context in which it occurs. Aspects such as background noise, social setting, network QoS (Quality of Service) and many others influence the user experience and must be considered when evaluating the usefulness of the service. As shown in e.g. [13] it is possible to simulate the context in which the service is envisioned to operate in the lab. However, doing so is not always trivial and the level of realism of such simulations is questionable.

Several comparative studies of field and laboratory testing use the number and type of uniquely identified usability errors as a measure for the effectiveness of the method. However, this might not always yield a true picture of the usefulness of a service. It could be argued that there are other aspects of usefulness which are of equal interest. Most importantly is of course if and how the service is used at all. Questions like when, where, how long, etc. users are interacting with a service can hardly be determined through a lab test. Also, long term habituation with a service cannot easily be investigated. Some issues will not be apparent before the user has used the service for a while and incorporated it into the daily routine. These questions can often only be investigated in longitudinal field trials with intended end users in their natural environment.

The big question, as it is put in [13] is whether it is worth the hassle to test mobile systems and services in the field compared to a lab setup.

It is our firm belief that field trials can supplement and contribute to a greater understanding of user behaviour and preferences. Furthermore, we believe that longitudinal field trials will yield information, which cannot be obtained through other methods. Given the well-known logistic and resource problems of field trials in general and longitudinal trials in particular, it is clear that there is a need for efficient and accurate tools to facilitate such test setups.

Our goal is therefore to identify how such studies can be conducted in an efficient and effective way, thus making it more attractive for researchers to explore mobile interaction in the field. We aim to both show that a lot can be gained from conducting large scale field trials with many users over long periods of time, and to develop new methods and tools for doing such studies more efficiently. In particular, we turn our attention to the questions of how to collect and analyse large amount of quantitative data and how this contribute to the understanding of user behaviour and the usefulness of the service. Going beyond the relatively simple usage statistics such as frequency of use and counting errors, the goal is to find out what, where, when, how and ultimately why the service is used.

The aim of this paper is to make a case for evaluating mobile services based (at least partly) on the analysis of quantitative data collected on mobile devices during longitudinal field trials. We will do this by:

- Presenting a mobile eHealth service for diabetes management which has been used as a case study.
- Presenting an implemented framework for automatic data capture on the mobile device and processing of this data into useful information about the given service.
- Presenting results and experiences from a field trial where a diabetic user has been using the service in his everyday life for three months.

It is important to note that we will not attempt to draw any general conclusions about the presented service. Rather we want to demonstrate that a lot of interesting aspects can be investigated by doing an unsupervised field test with automatic data logging and then analysing the pure quantitative data set.

1.2 Paper Outline

The following outlines the rest of the paper: Section 1.3 gives the background for the case study and 1.4 discusses related research. Section 2 describes the DiasNet Mobile prototype from a user's point of view. Section 3 presents the framework for data logging and analysis and details on the logging and processing of data. Section 4 describes the field trial setup and section 5 presents results from this. Section 6 discusses and concludes about these results and the use of quantitative log data. Finally conclusions and some thoughts on further work are given.

1.3 Study Background

The work presented in this paper is part of a larger study with the aim of showing how an existing decision support system for diabetes management and advisory can be integrated into a mobile and personal service for everyday use. This service has been coined DiasNet Mobile as it is based on the existing system called DiasNet.

DiasNet is a Danish service for type-1 diabetics. It is a web based interface for reporting and viewing data that are relevant to the management of diabetes i.e. blood glucose (BG) measurements, insulin (IN) injections and carbohydrate (CH) intake. It is the results of a research project conducted by The Medical Informatics Group at the Department of Health Science and Technology at Aalborg University [1]. The experiences from DiasNet (see [5] and [6]) forms the basis for the further development of the mobile service presented in this paper.

Optimal diabetes management is about keeping the BG level within defined thresholds. If the level gets too low, the user will get very uncomfortable as he/she experiences hypoglycemia and if it becomes too high, the user experiences hyperglycemia, which is the primary cause for almost all the long-term diabetic complications. To aid the user in this management, DiasNet uses the DIAS (Diabetes Advisory System) decision support system which is based on a CPN (Causal Probalistic Network) that models the human carbohydrate metabolism [5]. The system can predict the development of the user's blood glucose levels by a simulation based on previous BG measurements, when and what he/she eats and when, how much and what type of insulin is injected. The service will be described in more detail in section 2.

DiasNet Mobile has been developed during the MAGNET (My personal Adaptive Global NET) project and it's continuation project MAGNET Beyond [2]. MAGNET is focused on introducing novel technologies for personal networking with a profound emphasis on the user. User-centricity is thus a key element in the project and through the whole development process a UCD (User-centered design) approach has been taken by directly involving diabetics, medical doctors and nurses in the process. The current prototype is based on a user workshop, a usability evaluation of both a lo-fi and hi-fi prototype in the lab, and finally a field trial (see [10] and [11] for more details). The field trial presented in this study is thus part of the final evaluation of the service.

1.4 Related Work

Many methods have been developed for automating evaluation of usability/usefulness. Ivory and Hearst surveys many of these in [9] and categorizes them based on their proposed taxonomy for automated usability evaluation methods. Applying their taxonomy to the framework used in this study, the method class would be classified as "testing" with the methods being a combination of "log file analysis", "performance measurement" and "remote testing". The automa-

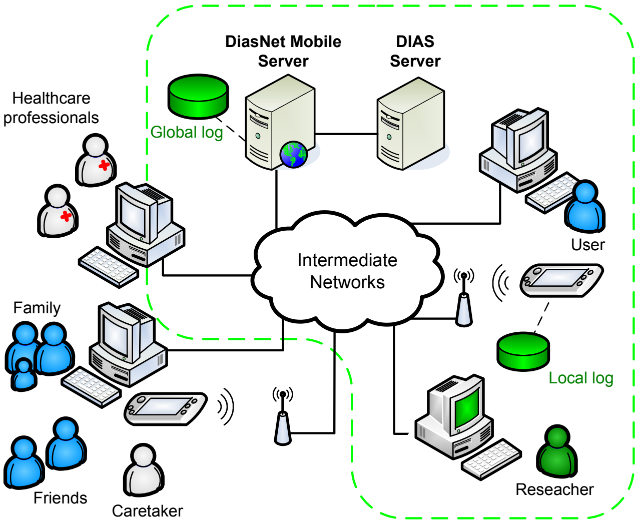

Figure 1: The architecture of the envisioned system. The dashed line marks the implemented prototype including the data logging framework used in the field trial.

tion type is both "capture" and "analysis", and the effort level is "model development" by the researcher and "informal use" of the service by the user.

Hilbert and Redmiles has surveyed a large amount of methods for extracting usability information from captured user interface events [8]. Considering the methods used in the framework presented in this paper in accordance with the categorization proposed in [8], they can be summarized as: transforming event streams, performing counts and summary statistics, visualization of the results for and to a lesser degree detecting sequences of events.

No systems or frameworks in either of these surveys are directly comparable to this study, and none are directly focused on mobile systems and services or field trials. More recently frameworks such as EDEM [7], MyExperience [4]and ContexPhone [15] has been created and used in studies somewhat similar to this one e.g. [3]. This study differs as it is looking more into the general usefulness and usage patterns at a higher level of abstraction.

2. THE SERVICE

An overview of the whole system is given in figure 1, which shows the overall architecture and intended user groups of the system. The dashed line marks what has been implemented in the current prototype. Also shown are the elements of the data logging framework which will be further elaborated in section 3. The system is integrated with the already existing DIAS server and thus the existing DiasNet web interface can be used both by the health care professionals and the user.

The field trial has focused entirely on the diabetic user and his interaction with the mobile part of the service. Primary devices are a Nokia 7710 Smartphone and a BG meter enhanced with a Bluetooth radio so that measurements are

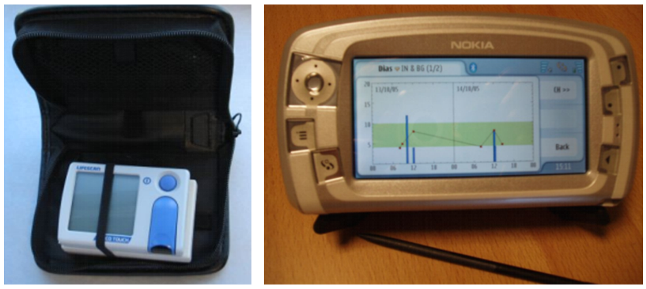

Figure 2: The devices used in the field trial: a Nokia 7710 and a modified BG-meter with Bluetooth.

directly sent to the phone. The devices are shown in figure 2.

Interacting with the service is done through an application running on the mobile phone. The interface is built as a hierarchy of screens and transitions as illustrated in figure 3. The functionality can be categorized in three main groups: "input of data", "viewing data" and "other functionality" e.g. settings.

Seen from a diabetes management point of view and from functionality of the DIAS system, some of these functionalities are more essential than other. These will in the following be referred to as the primary functionalities, and more specifically they are:

Blood glucose measurements (BG): Can be transferred either automatically from the BG meter or entered manually through the phone.

Insulin injections (IN): Manually entering how much long and short acting insulin is injected.

Carbohydrate intake (CH): Entering the consumption of carbohydrates e.g. after meals.

Viewing data: Can be done either by using the graph or table functionality. The graph view is where the DIAS simulations can be viewed.

Figure 3 shows the screens for the primary functionality. The results presented in this paper will focus on these.

3. DATA LOGGING FRAMEWORK

A framework for automatically logging and processing data for evaluation has been developed. As can be seen in figure 4 the framework consists of five main modules each representing essential phases:

Capture: The logging of events in the user interface is done by code instrumentation. During runtime events will be sent to a central data logger that stores it in local memory. It is a transparent process for the user and has virtually no influence on the performance of the application.

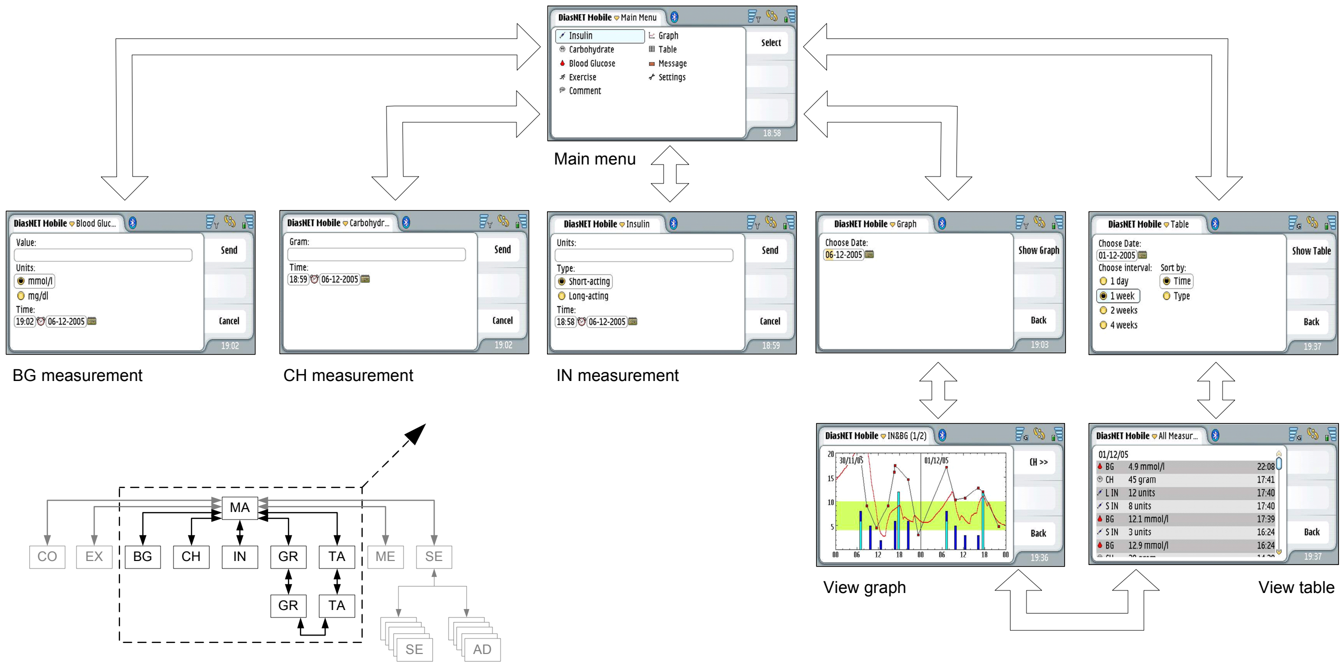

Figure 3: Overview of the GUI in DiasNet Mobile. The small diagram in the lower left corner shows the transitions diagram for the whole service. The main functionality is highlighted and shown in detail

Reporting: The reporting is done wireless over GPRS either automatically during shutdown of the service or manually by the user (e.g. after the trial).

Interpretation: The raw data log has to be stored and sent in a very compact manner to minimize the data to transfer over GPRS and the use of resources on the mobile device. Thus it has to be interpreted to be readable by the experimenter and the analysis module. The interpretation module is both a parser for the raw data log format, but it also serves as a preprocessing filter for handling potential errors. (i.e. temporal inconsistencies, out-of-bounds values and invalid log entries not conforming to the log protocol). As the logging is running autonomously such errors must be anticipated.

Analysis: This module is by far the most complex module. Building up a data structure for reasoning about the observed data, and to find both direct and derived statistics and facts about the usage. The logical structure of the data is elaborated in the end of this section.

Output: The results from the analysis must be transformed into useful data structures i.e. tables, graphs and diagrams for the researcher. Furthermore, the entire log is automatically annotated so that each entry is written out for human readability and annotated to get basic derived information such as duration and transitions.

The following are examples of what is logged:

- When the service is started and stopped
- UI events e.g. buttons pressed
- Screen transitions
- Frequency and duration spent in the various screens
- Any changes in settings
- Erroneous data entries, exceptions and any unexpected system behavior
- Use of the Bluetooth BG-meter

Since it is a medical service and real patients are involved no personal and sensitive data are logged, e.g. the exact values of BG or IN.

Intuitively it seems better to log too much than too little but the downside of an aggressive logging policy is information overflow. This necessitates processing of the data in order to figure out which data are "good" i.e. holds evidence to the things which are investigated in the experiment and to separate the noise.

The processed data set is the core information for evaluation of the system. An object-oriented representation of the data is automatically built up from the interpreted log, which can then be used to infer about the usage on a more general level.

The user's interaction with DiasNet mobile (and most other services) can be viewed as a series of sessions in which a number of activities are performed. The temporal aspect of usage is of great importance when trying to identify patterns of usage. So, time is added as another level in the hierarchy. The data is structured so that each day can have a number of sessions with a number of activities as shown in figure 4.

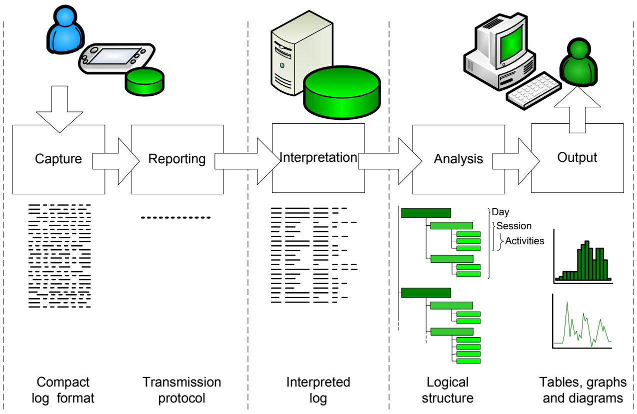

Figure 4: Framework for remote automated capture and analysis of usage data.

4. FIELD TRIAL SETUP

The field trial was designed as part of the development of DiasNet Mobile. The overall goal was to see what impact the service would have in the everyday life of diabetic patients by enabling them to manage their disease electronically anytime and anywhere. Thus it was necessary to evaluate the service with real diabetic users over a longer period of time in a natural setting.

The field trial has been divided into a pilot phase and a main phase. The pilot phase was run as a three month experiment with one user. Currently only the pilot phase has been conducted, and the experiences from the pilot study will be used to refine both the service and the main field experiment before the larger scale main phase is initiated. The main phase is further discussed in section 6.2.

The user selected for the pilot experiment was a middle aged male type-1 diabetic with 13 years of experience in disease management. He had already participated in the evaluation of the web based version of DiasNet and thus he was familiar with the concepts and terminology. He is a proficient user of mobile devices and services. The choice of an experienced user was based on the fact that his computer literacy and experience would make him more tolerant of minor technical errors which could (and did) occur due to the prototype state of the service.

The experiment was started by an introduction meeting, where the service and field trial was presented for the user. After ten days a follow-up meeting was held to get some feedback and ensure that everything was working. The field trial was concluded after three months, where the user stopped using the service. A debriefing interview was held with the user to get some feed-back on both the service and the field trial as such.

5. RESULTS

This section will present and discuss key results and observations from the field trial based on the data log. As stated earlier, these are not meant to support final conclusions about the usefulness of DiasNet Mobile, but rather to exemplify the possibilities for exploring different aspects of

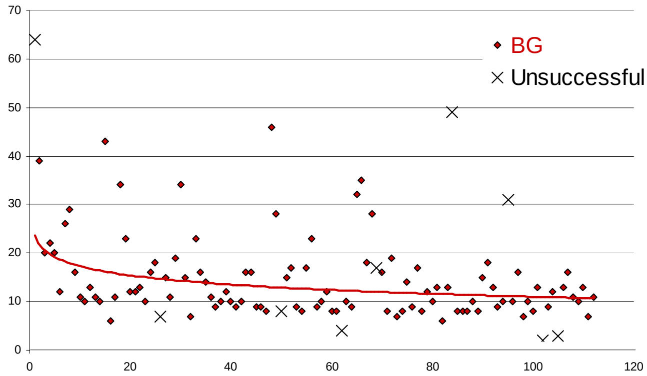

Figure 5: Graphs showing the usage of the blood glucose (BG) reporting functionality. The duration for each time the user tried to submit a measurement is plotted and a tendency line for the data set is shown.

usefulness by analyzing the quantitative log data.

The focus is on the primary functionality: reporting of blood glucose (BG), carbohydrate (CH) and insulin (IN) measurements and viewing the graph and table (View). The results are presented and discussed with regard to performance (section 5.1 and usage patterns (section 5.2) of the primary functionalities.

5.1 Performance

The performance of the main functionalities was investigated by plotting the duration for each time the user performed one of the corresponding activities. Figure 5 shows a plot of all the times the user has used the functionality for reporting blood glucose measurements (BG) on the phone. Several interesting things can be seen in this plot:

- The X's marks unsuccessful activities i.e. times where the user did not submit the value, but chose to cancel or shut-down the service. This has occurred relatively few times but these occurrences are quite equally distributed over the period.
- Although the plotting indicates that it mostly takes 10-15 seconds to enter and submit a value; there are still several times where considerably more time was used. The most extreme outliers are unsuccessful activities with either very high or too low duration, but numerous other times the user took 2-3 times the average.
- Even though the data is noisy, the plot still gives a good picture of the users performance with the given activity. The tendency line in figure 5 could be considered as an expression of the learning curve for the given functionality. In this case the activity is rather simple, but for larger and more complex tasks it might be possible to see a significant improvement over time.

The graphs for insulin (IN) and carbohydrates (CH) looks similar, and the same tendencies can be spotted. It does not

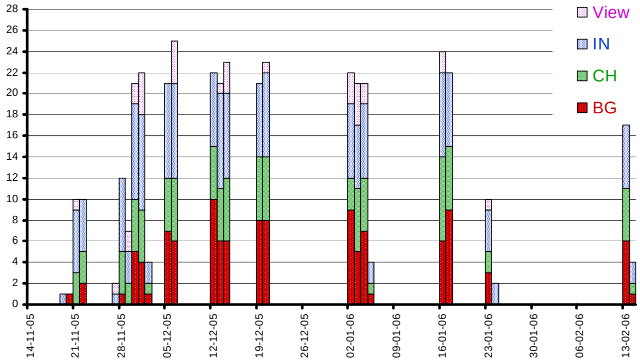

Figure 6: The usage of the primary functionalities accumulated for each day over the whole period.

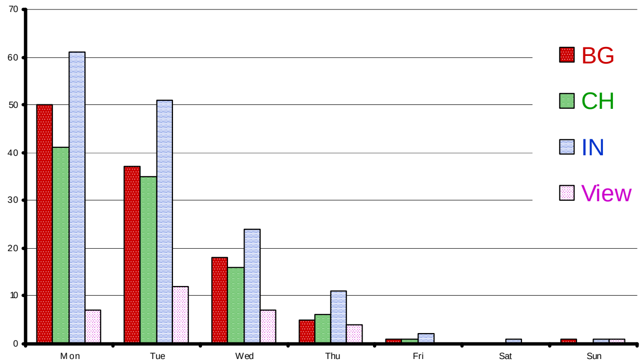

Figure 7: The usage of the primary functionalities on each weekday accumulated over the whole period.

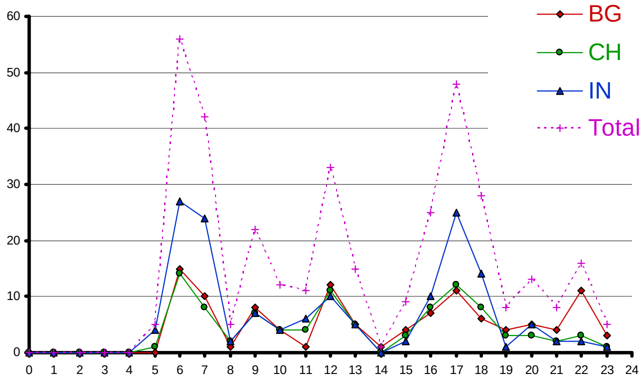

Figure 8: Usage of the primary functionalities based on time of day accumulated over the whole period.

make sense to look at the performance of the graph and table view in this way, since the time spent it is very dependent on the what the user is looking for.

5.2 Usage Patterns

When looking at usage patterns, the temporal dimension is very important. In the following the usage of the service is explored at various granularities of time.

Figure 6 shows the usage of primary functionalities over the whole period as a bar chart. Several observations can be made from this overview:

- In the beginning the BG functionality is not used much, but for the rest of the period it seems like the three primary input functionalities are used consistently and almost equally. The reason for the lack of BG activity in the beginning is due to the fact that the user used automatic BG reporting from the Bluetooth enabled BG-meter. Thus he did not have to enter the BG values in the system manually. Due to technical problems, the BG meter stopped working after a couple of weeks, and it was decided not to use it for the rest of the test.
- During most of the period the usage is concentrated in small chunks of three to four consecutive days.

Figure 7 shows have the use of primary functionally looks when the weekdays are considered. Again some observations can be made:

- It can be seen that the 3-4 four day chunks are usually in the beginning of the week. This has both to do with the fact that the user does not normally measure his BG levels in the weekends, as he considers this his free time. His eating habits are quite different during this time, and thus he does not want to feel that he is under constant observation. Looking closely at the dates in figure 6 it can also be seen that there are no entries during the Christmas period and a week late in January (probably winter vacation). Otherwise the use has been systematic and consistent throughout the trial period. This fits with the reluctance to use the equipment in weekends.

To see if there is a specific pattern of use during the day, the use of primary functionalities has been accumulated and plotted as a graph over the range of hours in a day. Figure 8 shows this. Some observations are:

- Peaks can clearly be seen in the morning at breakfast (6-7), at lunch (12) and dinner time (17-18).
- It is very rarely used around 8 and 14 (and never during the night).

Usage patterns can also be explored at session level. Table 1 shows a combinatory matrix of the primary input functionalities and how often each was used in the same session as the other. Some observations can be made table 1:

- There are many sessions with multiple inputs. When considering the 38 sessions without any input, it should be noted that 13 of these had a duration of less then 5 seconds, indicating that the services was unintentionally started and can be omitted. Also it shows only when at least one measurement of the given types has

		BG			
		TRUE		FALSE	
		IN		IN	
		TRUE	FALSE	TRUE	FALSE
CH	TRUE	65	6	18	8
	FALSE	8	23	8	38

Table 1: Combinatory matrix for when more than one of the primary input functionalities were used in the same session.

been reported. A large proportion of the sessions had multiple inputs of at least on type. These findings suggest that the service can be made more efficient by allowing the user to bundle inputs of different types. This feature was removed during the design phase for simplicity of the interface, but should be reconsidered based on the field trial results.

6. DISCUSSION

As the results presented in the previous section are based on a single user it is impossible to generalize and conclude on the service based on this. However, our intention here is not to make general conclusions about the usefulness of the particular service, but rather to use it to illustrate the points made in section 1 about longitudinal testing and quantitative logging.

Follow-up and debriefing interviews with the test participant confirmed many of the findings and formed an important part of the trial. Obviously, the test user(s) will be able to provide highly qualified in-depth responses to the service, compared to the responses which could be expected after a lab-based test of limited duration.

One quite surprising finding was that the user did not use the graph functionality very often, even though this was thought to be one of the strongest and most useful features of the service. Users that participated in the usability test performed in the lab had all liked it very much. The reasons for this are probably that the user is an experienced diabetic, and knows his blood glucose metabolism well. Thus it has no utility to him. Also he knows the graphs from his experience with the DiasNet web service and it thus has no novelty value to him.

The problem with lab tests is that users might say, that they will be interested in using a given feature after trying it a couple of times, but in reality they might scrap it after a few weeks or find some workarounds to do the same task in a more convenient way. Not because they are intentionally lying, but simply because they cannot imagine how it would be to use the given feature in the long run.

A field trial with data logging like the one presented in this paper is not very likely to be enough for evaluating the usefulness of the service. For getting the user's opinions and feelings about the system, more qualitative oriented methods are needed as well. Combining a longitudinal study with methods like user interviews and questionnaires is believed to be a feasible approach, since they naturally complement each other and are easily combined. The data from the study can inform the researcher and thus the questions asked can more precisely target problematic and interesting aspects. In this regard the quantitative basis might actually improve the quality of data gathered by the qualitative methods. If log data is continuously submitted from the users' devices during the whole period, it will allow the researcher to inspect preliminary data at an early state and spot trends and tendencies to include in the qualitative data collection.

As both the data capture and analysis process has been automated to a large degree, it will require minimal effort to produce the same data for the main phase of the field trial even though it will involve more users. This means that the researcher can put most of the focus and energy into the interpretation of data and using this to critique and improve the service.

It could be argued that the data set could be divided into two parts based on how much and when the user has used the service: one for the intro phase where the user gets accustomed to the interface and functionality and a part which represents the "natural" behavior of the user. In this study the whole data set is used as one, as it was not possible to clearly separate the initial experimentation from the normal usage.

One of the big limitations in this type of unsupervised experiments is, that it is impossible to explain outlying values like the ones in figure 5. It seems reasonable to assume that the user might have been interrupted during use or some other external influence caused him to use much more time on the simple task, but nothing can be said for sure.

There are other inherent problems with the nature of the data produced by this sort of experiment due the lack of control and direct monitoring. Large amounts of noise may be introduced in the dataset, which makes the analysis more difficult. E.g. one particular day the user has systematically browsed through all the functionality. This is hardly representative of normal use and thus it should not be included in the statistics. If enough data was collected, such noise would however be smoothed out. This raises the question of how much data is needed, and how to determine what is noise and what is not from a pure quantitative data set. This is a very interesting problem. We believe this can to a very large degree be automated using sophisticated statistical pattern recognition methods.

6.1 Conclusions

The experiences from this study has shown that a lot can be learned about how a service is used by looking at the quantitative data from a longitudinal field trial, and that this can be tracked back to inform the design of the service. Even considering the limited scope of the conducted field trial, several observations have been made that could not have been found in the lab. Some of these are even in contrast with findings from lab evaluations earlier in the development phase. This indicates that something can indeed be gained by evaluating mobile services in the field. Also, by studying the logging data and subsequently interviewing the test person(s) about specific observations the inherent weakness of unsupervised field trials can be amended.

The framework for automated logging and analysis has proven to be a good solution for acquiring and processing raw usage data into a useful representation on which to base an evaluation - especially with regard to the utility side of usefulness. Due to its autonomy and scalability properties the framework is well suited for longitudinal studies with many users over long periods.

The downside of this approach is that a working prototype is needed, and thus the method will only be applicable in late phase of development. Also the framework needs to be developed. If some general automated method based on a further formalized way of representing usage data can be created, then it might prove to be a feasible and cost-effective way of evaluating hi-fi prototypes of mobile services.

In conclusion, the use of quantitative usage data for evaluating the usefulness of DiasNet Mobile has proven to be gainful, and it is believed that the type of results presented in this paper could also be attained if a similar experiment was conducted on another mobile service.

6.2 Further Work

As the final evaluation of the DiasNet Mobile service, the planned main phase of the field trial with 7-10 users over three months will be conducted. This study will be done with existing DiasNet users and/or newly diagnosed diabetics. By having more users it is believed that enough evidence can be collected to make solid conclusions about the usefulness of the service.

Also, further development of the logging and analysis framework will be done in order to make it into a general tool. It is also planned to try to include contextual information, to see if this can be used to make better inferences from the quantitative usage data.

Acknowledgements

The development of DiasNet Mobile has been part of MAGNET Beyond which is a continuation of the MAGNET project (www.ist-magnet.org). MAGNET Beyond is a worldwide R&D project within Mobile and Wireless Systems and Platforms Beyond 3G. MAGNET Beyond will introduce new technologies, systems, and applications that are at the same time user-centric and secure. MAGNET Beyond will develop user-centric business model concepts for secure Personal Networks in multi-network, multi-device, and multi-user environments. MAGNET Beyond has 32 partners from 15 countries, among these highly influential Industrial Partners, Universities, Research Centers, and SMEs.

7. REFERENCES

[1] URL: http://www.mi.hst.aau.dk/˜spp/.

[2] URL: http://www.ist-magnet.org.

[3] N. Eagle and A. Pentland. Reality mining: Sensing complex social systems. *Personal and Ubiquitous Computing*, 10(4), 2006.

[4] J. Froehlich, M. Chen, S. Consolvo, B. Harrison, and J. Landay. Myexperience: A system for in situ tracing and capturing of user feedback on mobile phones. In *MobiSys*, 2007.

[5] O. Hejlesen, S. Andreassen, R. Hovorka, and D. Cavan. Dias - the diabetes advisory system: An outline of the system and the evaluation results obtained so far. In *Comp. Meth. Prog. Biomed*, 1997.

[6] O. Hejlesen, S. Plougmann, and D. Cavan. Diasnet - an internet tool for communication and education in diabetes. 2000.

[7] D. M. Hilbert and D. F. Redmiles. An approach to large-scale collection of application usage data over the internet. In *ICSE '98: Proceedings of the 20th international conference on Software engineering*, pages 136–145, Washington, DC, USA, 1998. IEEE Computer Society.

[8] D. M. Hilbert and D. F. Redmiles. Extracting usability information from user interface events. *ACM Comput. Surv.*, 32(4):384–421, 2000.

[9] M. Y. Ivory and M. A. Hearst. The state of the art in automating usability evaluation of user interfaces. *ACM Compututing Surveys*, 33(4):470–516, 2001.

[10] K. L. Jensen, C. F. Pedersen, and L. B. Larsen. Towards useful and usable services in personal networks. In *First International Workshop on Personalized Networks*, 2006.

[11] K. L. Jensen, C. F. Pedersen, and L. B. Larsen. Diasnet mobile: A personalized mobile diabetes management and advisory service. In *Second Workshop on Personalization for E-Health*, 2007.

[12] J. Kjeldskov and C. Graham. A review of mobile hci research methods. In *Mobile HCI*, pages 317–335, 2003.

[13] J. Kjeldskov, M. B. Skov, B. S. Als, and R. T. Høegh. Is it worth the hassle? exploring the added value of evaluating the usability of context-aware mobile systems in the field. In *Proceedings of the 6th International Mobile HCI 2004 conference*, pages 61–73, Berlin, 2004. Springer-Verlag.

[14] J. Nielsen. *Usability Engineering*. Morgan Kaufmann, 1993.

[15] M. Raento, A. Oulasvirta, R. Petit, and H. Toivonen. Contextphone - a prototyping platform for context-aware mobile applications. *IEEE Pervasive Computing*, 4(2), 2005.