



ELSEVIER

journal homepage: www.intl.elsevierhealth.com/journals/ijmi

Risk analysis of information security in a mobile instant messaging and presence system for healthcare

Erlend Bønes^a, Per Hasvold^a, Eva Henriksen^{a,*}, Thomas Strandenæs^b

^a Norwegian Centre for Telemedicine, University Hospital of North Norway, P.O. Box 35, NO-9038 Tromsø, Norway

^b Well Diagnostics AS, P.O. Box 6431, NO-9294 Tromsø, Norway

ARTICLE INFO

Article history:

Received 9 February 2005

Received in revised form 8 June 2006

Accepted 8 June 2006

Keywords:

Instant messaging

Mobility

Healthcare

Information security

Risk analysis

ABSTRACT

Introduction: Instant messaging (IM) is suited for immediate communication because messages are delivered almost in real time. Results from studies of IM use in enterprise work settings make us believe that IM based services may prove useful also within the healthcare sector. However, today's public instant messaging services do not have the level of information security required for adoption of IM in healthcare. We proposed MedIMob, our own architecture for a secure enterprise IM service for use in healthcare. MedIMob supports IM clients on mobile devices in addition to desktop based clients.

Methods: Security threats were identified in a risk analysis of the MedIMob architecture. The risk analysis process consists of context identification, threat identification, analysis of consequences and likelihood, risk evaluation, and proposals for risk treatment.

Results: The risk analysis revealed a number of potential threats to the information security of a service like this. Many of the identified threats are general when dealing with mobile devices and sensitive data; others are threats which are more specific to our service and architecture. Individual threats identified in the risks analysis are discussed and possible counter measures presented.

Discussion: The risk analysis showed that most of the proposed risk treatment measures must be implemented to obtain an acceptable risk level; among others blocking much of the additional functionality of the smartphone. To conclude on the usefulness of this IM service, it will be evaluated in a trial study of the human-computer interaction. Further work also includes an improved design of the proposed MedIMob architecture.

© 2006 Elsevier Ireland Ltd. All rights reserved.

1. Introduction

Use of instant messaging services is becoming increasingly popular with Internet based systems like America Online's Instant Messaging, AIM (<http://www.aim.com/>), Microsoft's MSN Messenger (<http://messenger.msn.com/>), Yahoo! Messenger (<http://messenger.yahoo.com/>), and ICQ (<http://www.icq.com/>).

However, public instant messaging systems have been criticised for having a number of security weaknesses [1–3]. These weaknesses include the facts that the IM clients are always on, that logs can contain sensitive information, and that the communication goes via an externally controlled server. Most IM services were never intended for secure communication in the first place [2]. The rapid growth in the number of public IM users has created a new security concern for IT managers.

* Corresponding author. Tel.: +47 95731836; fax: +47 77754098.

E-mail address: eva.henriksen@telemed.no (E. Henriksen).

1386-5056/\$ – see front matter © 2006 Elsevier Ireland Ltd. All rights reserved.

doi:10.1016/j.ijmedinf.2006.06.002

New worms and viruses are increasingly using IM to spread, and 5–10% of the IM traffic today can be categorised as spam over IM (SPIM) [4].

Within the healthcare sector information security aspects are of vital importance, and may be of serious hindrance for the adoption of IM based services. In this paper we will examine the feasibility of using instant messaging systems in the healthcare sector from the viewpoint of information security.

Healthcare professionals are working in a mobile environment with rapid changes in their availability status, and they are exposed to interruptions at any time, anywhere. In addition to traditional desktop IM clients, IM for use in healthcare settings should therefore also offer clients on mobile devices.

In order to take care of both mobility and security aspects, we have proposed our own architecture: the MedIMob system. An overview of the MedIMob architecture is presented in this paper. Components of the MedIMob system have been further developed at the Norwegian Centre for Telemedicine (NST).

The main contribution of the paper is the results from a risk analysis of the MedIMob system, based on the architectural design of the system. The results of this risk analysis may be valid to other systems with a similar approach. In the risk analysis the assumed environment for the system was a hospital department, and communication within the department and between IM clients inside the department and IM clients outside. Information security challenges were identified as a number of security threats of different risk levels. Solutions are proposed for improvements of the unacceptable threats.

2. Background

Instant messaging (IM) is a lightweight near-synchronous communication technology. Technically it offers asynchronous communication, but it is used as synchronous communication because the messages are delivered almost in real time. Additional functionality for publishing and subscribing to presence information makes it possible for the users to see which other users and resources are available at any time. Presence information can be based on, e.g. schedules and calendar information, user settings, or keyboard activity.

2.1. Instant messaging use in workplace and healthcare

IM has proven its value as media for personal communication, both for peer-to-peer channels and as a conferencing platform. This adoption has mainly been seen in private social contexts. Adoption in workplace has been slower; one reason could be apprehensions that the service would be used primarily for private purposes (chatting) resulting in misuse of time and resources; another reason could be the informality of the service and the lack of security and documentation.

Based on early results from studies of IM use in enterprise work settings [5–8], we believe that IM based services may prove useful also within the healthcare sector. There are a few descriptions of IM use within healthcare. One of them is the EU-funded research project PICNIC from the fifth framework

programme (FP5) of Information Society Technology (IST). The project describes IM as one of several collaboration components in a healthcare network [9]. In their system IM is used to discover the online presence of connected experts that can be invited to collaborate, e.g. to give a second-opinion on a medical case. The invited expert uses IM to confirm that a second-opinion can be given or to request additional information.

IM solutions for enterprises, including healthcare, have been developed by different companies. One of the most interesting approaches is offered by UnBound Technologies Inc. [10], where IM is part of a business process management system for hospitals. This is a presence-based enterprise messaging platform that enables wireless communication with two-way alerting and notifications. The notifications can come not only from co-workers but also from legacy systems in, e.g. laboratory and radiology.

2.2. Characteristics of the healthcare environment

Healthcare professionals are working in a mobile environment with rapid change in their availability status. A Danish study [11] which focuses on local mobility in a hospital department, divides the need for mobility into four categories: (1) the need for being at different physical *places*, (2) the need to access *knowledge*, (3) the need to use shared *resources*, and (4) the need to get in contact with specific *persons*. Healthcare workers are exposed to interruptions at any time, anywhere. They usually carry pagers which give them a phone number to call, but without additional information about the reason for the request. Thus, it is difficult for them to decide the importance of the interruption.

In addition to the mobility, there are also other aspects that make the healthcare domain different from other arenas. One aspect is the need for documentation and traceability of decisions and actions. Another aspect is the high need for information security mechanisms caused by the privacy requirements related to communication of sensitive patient identifiable information. Confidentiality requirements originate from the professional secrecy and non-disclosure agreement imposed to all healthcare workers. Requirements to electronic communication of patient information come from national legislation in European countries, based on EU's regulation on processing of personal data (95/46/EC) from 1995 [12], and from the American Health Insurance Portability and Accountability Act of 1996 (HIPAA) [13]. At the lowest level these requirements become apparent through the security policies of the affected organisation.

2.3. Risk analysis of information security

Security risk analysis is a basic requirement of ISO 17799, internationally recognized as the generic information security standard [14]. Risk analysis is also required by national legislation as a vital part of an information security management system for any organisation. Risk analysis is performed with respect to the main information security aspects confidentiality, integrity and availability. The risk acceptance criteria are defined by the information security policies of the affected organisation.

There are many methods and guidelines for how to perform risk analysis, but all of them include the central tasks of

- identifying the threats or possible unwanted incidents,
- analysing the impacts and probabilities of these threats,
- evaluate risks with respect to the acceptance criteria.

Our experience is based on the EU-funded research project CORAS from the fifth framework programme (FP5) of Information Society Technology (IST) [15,16] where a methodology for risk analysis was developed and tested on e-health systems. The methodology was based on the Australian and New Zealand standard for risk management (AS/NZS 4360/1999) [17], which clearly sets out the risk analysis process in five main steps:

1. Context identification: a description of the subject for analysis, i.e. the analysed system and its environment.
2. Threat identification: identify what could possibly happen.
3. Impact and probability analysis: a consideration of the consequences of the threats and the likelihood that these consequences may occur.
4. Risk evaluation: relating the resulting risk level with risk acceptance criteria.
5. Risk treatment: identification and assessment of treatment options.

3. Architecture of the MediMob system

To study the information security properties of IM we devised a preliminary architecture for an enterprise IM which embeds a number of the information security techniques usually deployed in areas with high security requirements. This architecture served as basis for the risk analysis presented later in the paper.

In our architecture we propose to use instant messaging and presence techniques to handle the availability and presence aspects, with mobile clients to support the mobil-

ity aspects of healthcare workers. With mobile smart phones instead of pagers, the healthcare professional is able to see what the request is about and based on that decide whether it is necessary to respond immediately or after a while. The use of mobile phones also makes the service reachable outside the healthcare institutions, by communication via GPRS (general packet radio service).

Work on a prototype for an experimental Jabber/XMPP based IM system which includes both desktop and mobile clients, has been going on at the Norwegian Centre for Telemedicine. Our system uses an open source Jabber/XMPP server and an open source desktop client. On the mobile side, a prototype client is being developed for the Java MIDP 2.0-platform which will be deployed on the Sony Ericsson P900 smartphone with Symbian OS. A smartphone is a combination of a traditional mobile phone and a PDA (personal digital assistant).

The proposed architecture is based on the extensible messaging and presence protocol (XMPP) as described by the IETF working group for XMPP [18,19]. Our modifications to the XMPP architecture are related to the organization of the XMPP server's handling of the traffic between clients on the Internet and clients located in secure or private networks.

Our system is intended for use within healthcare, where some of the users are in a secure network zone inside a hospital department. They are using desktop clients connected to the internal network or mobile clients connected through a wireless local network (WLAN). Users outside the secure network zone can connect from mobile clients on smart phones via public networks (GPRS), and they have to go via the Internet to communicate with clients in the secure zone.

We place no restrictions on which type of information the users can exchange using the IM service, and this can obviously be patient identifiable sensitive information. Our opinion is that the service will mainly be used for short messages and questions/responses, in the same way as the telephone would be used for these needs. This is information that would not naturally go directly into the patient's health record.

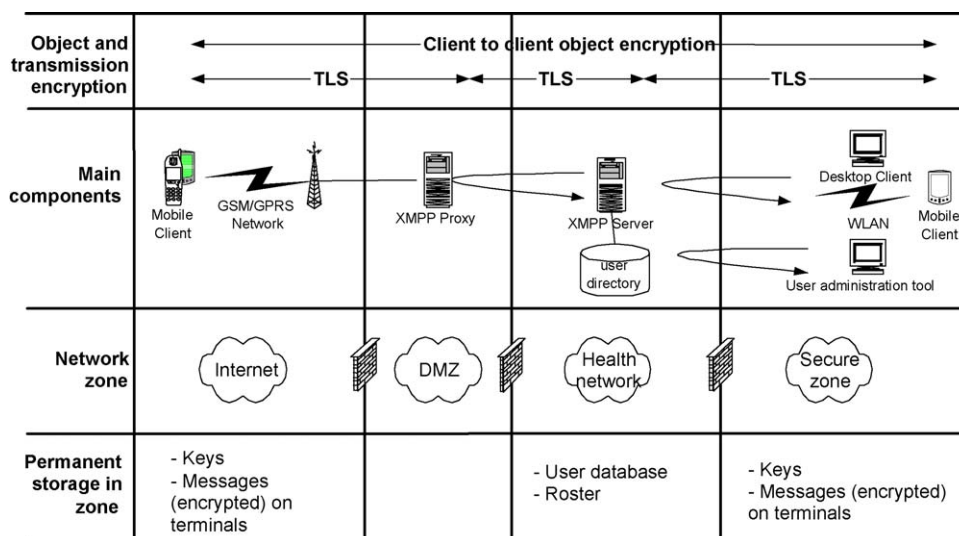


Fig. 1 – MediMob system architecture.

To ensure confidentiality requirements, the communication between the clients is protected by end-to-end encryption at application level. Additional security mechanisms are included, based on the solutions implemented in a system for communicating sensitive information between insecure and secure networks, developed by NST [20]. As shown in Fig. 1, the network used by healthcare workers is separated into different zones with corresponding security levels: the secure zone, the internal zone, a demilitarized zone (DMZ), and the open Internet. The local network in the hospital is regarded a *secure* zone. It is separated from the healthcare network by firewall solutions. The Norwegian healthcare network connects a large number of healthcare institutions. It is in this case regarded an *internal* zone. The healthcare network is separated from the open Internet by a double set of firewall solutions. The DMZ is at the interface between the healthcare network and the Internet. Requirements from the Norwegian Data Inspectorate state that communication cannot be initiated from a less secure network to a more secure network or zone. Use of polling between network zones is a way to satisfy these requirements.

The architecture illustrated in Fig. 1 is summarised by the following points:

- The XMPP server is located in the healthcare network. All communication with external clients goes through an XMPP proxy that is placed in the DMZ of the healthcare network. This is a “reverse polling proxy”, where the only communication between the server and the proxy happens when the server polls the proxy for new messages. This makes the server less exposed to traffic from the Internet.
- A user directory, accessed by the XMPP server via LDAP (lightweight directory access protocol), contains subscriber information such as users’ addresses, status, and their certificates. This directory can only be accessed from the XMPP server and the administrator consol.
- Communication between the XMPP server and the clients in the secure zone is also based on polling. Data from the XMPP server is only sent to clients in the secure zone as a result of a poll.
- Point-to-point communication is secured with SSL/TLS (secure socket layer/transport layer security). This implies a decryption and a new encryption at each node.
- End-to-end (client-to-client) encryption is implemented at application level. This is hybrid encryption, combining symmetric and asymmetric (PKI-based) encryption algorithms. The message itself is encrypted with a one-time session key (AES algorithm), and the session key is encrypted with the recipient’s public key (RSA algorithm) and delivered together with the message. This is in accordance with the proposals for the XMPP protocol [21].
- The exchanged messages shall not contain any executable code, only information types like text, images and sound. This precaution is meant to reduce the likelihood that messages can be carriers of malicious code.

Our view is that messages exchanged through an IM based conversation are of a transient nature, similar to a telephone conversation. Therefore, it has not been intended to let the MedIMob system automatically archive messages

containing patient information, for instance into the electronic health record. However, healthcare professionals are expected to indicate in the patient’s health record any relevant information related to this communication, like they would do with a telephone communication. Some sort of logging of the communication will also take place. Logging should include information about the sender, recipient and time of communication. The actual content of the message might also be logged, but in that case it should be encrypted.

4. Risk analysis method

To analyse the security challenges of an IM service for healthcare, we performed a qualitative risk analysis of the information security aspects of our proposed architecture and the intended environment. The goal was to identify security threats to the use of our instant messaging service within a hospital department, and find acceptable solutions to the threats.

Based on our experience from the CORAS project [15,16], we performed the risk analysis by going through the five main steps described in the Australian and New Zealand standard for risk management [17]:

1. The context identification was given by the architectural design of the MedIMob system and a description of the systems intended environment.
2. The threat identification was performed as a structured brainstorming between the project members and the discussion was summarised in a risk table with the following columns:

- unique ID of threat (threat number),
- description of threat or unwanted incidence,
- consequence value (and additional description, if any),
- likelihood value (and additional description, if any),
- risk value (as a product of consequence and likelihood),
- any other comments (including ideas for risk treatment).

In the structured brainstorming process a walk-through of the architecture was performed, using predefined guide-words and attributes. Guidewords were related to the security aspects confidentiality, integrity and availability, and to attributes like “internal” and “external” (threats), and “deliberate” and “accidental” (actions).

The risk table is non-static and is used as a tool throughout the whole process. During the brainstorming, all possible threats were written into the table, together with any relevant comments, also any comments related to consequence and likelihood. Afterwards, a clean-up of the table was performed, by grouping related threats or putting threats into a relevant sequence. At this stage each threat was given its unique ID (values for consequence, likelihood, and risk were added later on in the process).

This is a HazOp-like method (HazOp—hazard and operability study) for identification of threats that can lead to potential hazard. HazOp evolved as a method for analysing safety of systems in the process industry. It has later been applied to other areas and was used with success as one of several methods in the CORAS project [15,16].

Table 1 – Our definitions of values for consequence, likelihood, and risk level, used in the risk analysis of the MedIMob architecture

Consequence	
Very small	Does not affect confidentiality, integrity and availability of information. Can in some cases influence reliability for some users
Small	Short interruptions of availability for some users or groups of users. No breach of confidentiality or integrity
Medium	Interruptions of availability for all users or a group of users for a longer time period. No breach of confidentiality or integrity
High	Breaches of information confidentiality, integrity and availability which affect individual users, not the service as a whole
Very high	Breaches of information confidentiality, integrity and availability which affect all users and the service as a whole
Likelihood	
Very small	Very rare. Occurs less frequent than 0.1% of the time/cases
Small	Rare. Occurs between 0.1 and 1% of the time/cases
Medium	May happen. Occurs between 1 and 5% of the time/cases
High	Quite often. Occurs between 5 and 20% of the time/cases
Very high	Very often. Occurs more frequent than 20% of the time/cases
Risk level	
Insignificant	Acceptable
Low	Acceptable risk. The service can be used with the identified threats, but the threats must be observed to discover changes that could raise the risk level
Moderate	Can for this service be an acceptable risk, but for each case it should be considered whether necessary measures have been implemented
High	Not acceptable risk. Cannot start using the service before risk reducing treatment has been implemented

3. The identified threats and unwanted incidents were analysed. For each possible threat we evaluated its impact or consequence and the likelihood that it would occur. Each threat was given qualitative values for consequence

and likelihood (i.e. very small, small, medium, high, very high). Definitions for the qualitative values for consequence and likelihood were discussed and agreed on before the brainstorming. Our value definitions are presented in Table 1.

- The risk value for each threat was calculated as the product of consequence and likelihood, illustrated in a two-dimensional matrix. The unique ID of the threat was written into the corresponding cell of the matrix, as illustrated in Fig. 2. The shading of the matrix visualizes the different risk levels. Based on the acceptance criteria, we defined four risk levels: insignificant, low, moderate, high (see our definitions in Table 1). The risk level high was decided to be unacceptable. Any threat obtaining this risk level must be treated in order to have its risk reduced to an acceptable level. The resulting risk levels for individual threats were compared with the risk acceptance criteria.
- For all threats with a non-acceptable risk level, risk-reducing treatment was proposed and discussed.

5. Risk analysis results

Table 2 shows the threats that were identified during the risk analysis. Fig. 2 shows for each threat the estimated likelihood and consequences.

Many of the identified threats are general when dealing with mobile devices and sensitive data (threat ID 1–10 in Fig. 2); others are threats which are more specific to our application and architecture.

In our risk analysis we found five threats which had an unacceptably high risk level, as can be seen from the risk matrix in Fig. 2. Three of these threats (threats 8, 25, and 26) were related to the possibility that the mobile device is connected to the secure local network at the same time as it is connected to the Internet via GPRS (e.g. during synchronisation) or to another device via Bluetooth or infrared connections. This in turn causes possibilities for malicious attacks from external sources with corresponding threats to all security aspects.

In the following paragraph also threats with medium risk level are discussed. These may become more severe and they should as far as possible be treated (the discussion refers to the threat IDs in Table 2 and in the matrix, Fig. 2).

Consequence \ Likelihood	Consequence				
	1. Very small	2. Small	3. Medium	4. High	5. Very high
1. Very small		31		7	
2. Small	21		29, 30	9, 23, 24	4
3. Medium	11	27	14, 28	1, 3, 12, 15, 16, 18, 20	
4. High	13		2, 6, 17	5, 8	
5. Very high			19	25, 26	

Fig. 2 – Risk matrix for the MedIMob architecture, showing the diffusion of identified threats. Shading illustrates different risk levels (insignificant, low, moderate, high).

Table 2 – Threats identified during risk assessment

ID	Threats, unwanted incidents
1	Mobile device with an <i>active</i> client (logged on and unlocked) can be lost and/or grabbed by unauthorized persons who can pretend to be the logged on user and therefore send messages, receive/read messages, change presence information
2	Mobile device which is <i>passive</i> (turned off or locked as a result of a timeout) can be lost or grabbed by unauthorized persons The owner has no longer a device available
3	Mobile device which is <i>passive</i> (turned off or locked as a result of a timeout) can be lost and/or grabbed by unauthorized persons Who can try to get access to the unit by breaking the PIN-code or other methods (we assume that breaking the PIN-code results in a reset of the unit with loss of content in volatile memory)
4	Mobile device which is <i>passive</i> (turned off or locked as a result of a timeout) can be lost and/or grabbed by unauthorized persons Who attempt to break the client password to access the service (the PIN-code or any other locking mechanisms are assumed to be removed first)
5	Loss of external memory card (from the memory slot of the device)
6	Sensitive information in the memory card can be accessed by unauthorized persons (break of confidentiality)
7	Loss of external memory card (from the memory slot of the device)
8	Information in the memory card is lost and unavailable
9	Virus received from the PC during synchronisation (or from other services or networks, see below). Virus can make all sorts of damage to the device
10	Connection via Bluetooth or other services and networks that delivers data from outside (SMS, MMS, infrared). Could give full access to services on the device. Attack methods could be memory overwrite, exploitation of weaknesses in protocols, etc., services that can change the configurations of the device (SMS)
11	Unauthorized persons who find/grab the mobile device can reconfigure it, e.g. activate Bluetooth. This could give unauthorized persons access to services on the device or exploit weaknesses, see threat 8
12	Avoiding the internal security policy/firewall by unintentional bridging of internal network (via WiFi) and external network (via GSM/GPRS). Opens for attacks from the external network, and unfiltered traffic from the internal to the external network
13	The mobile device is always connected to the Internet (via GPRS), and therefore exposed to all kinds of attacks from the Internet (see threat 7)
14	For example: denial of service attack
15	The mobile device is always connected to the Internet (via GPRS), and therefore exposed to all kinds of attacks from the Internet (see threat 7)
16	For example: unauthorized persons can pose as the logged in person, send messages, receive/read messages, changes presence information. That is, break of confidentiality, integrity, availability
17	The XMPP proxy can be compromised and manipulated by unauthorized persons. It is placed in the DMZ (less likely that the server is attacked because it is in a more secure environment)
18	The proxy or the server environment is compromised in a way that makes the service unavailable or unreliable
19	The XMPP proxy can be compromised and manipulated by unauthorized persons. It is placed in the DMZ (less likely that the server is attacked because it is in a more secure environment)
20	Messages are not forwarded, or forwarded at the wrong time
21	The XMPP proxy can be compromised and manipulated by unauthorized persons. It is placed in the DMZ (less likely that the server is attacked because it is in a more secure environment)
22	Sender information can be modified. This can affect the integrity of the message, without being detected
23	The XMPP proxy can be compromised and manipulated by unauthorized persons. It is placed in the DMZ (less likely that the server is attacked because it is in a more secure environment)
24	Receiver address can be changed, break of confidentiality may occur
25	External or hostile client establish a connection to the service, either through proxy (from the Internet) or from the healthcare network, and authenticates itself with a valid username and password. Gets access to updated presence and subscription status on the server (can be viewed as one message) and similar services. In this way unauthorized persons can get access to sensitive information (e.g. the fact that one are having a communication with a psychologist), and make changes to roster (subscription information)
26	Sensitive data exists in clear text in the memory of the mobile device (for threat 1 and 2). Data is compromised, i.e. break of the confidentiality, if the mobile device is used by unauthorized persons
27	Loss of content of volatile memory, as a result of power outage, device failure, software error, etc. Received messages which are still unread can be deleted
28	Disclosure of encryption key stored in the mobile devices
29	If all communication shall be compromised from a lost mobile device, the central unit (XMPP-server) must also be compromised, where the SSL-communication is decrypted
30	LDAP-server is compromised
31	Can change the user's password for access to the server, resulting in denial of service
32	LDAP-server is compromised
33	Could try to register new uses. To do so, administration software has to be available to install the client (including installation of valid encryption key, which also requires a password to encrypt the shared key)
34	Stationary (desktop) client in secure zone is compromised, resulting in break of confidentiality, integrity, availability
35	The wireless network (WiFi) is compromised, resulting in break of confidentiality, integrity, availability

Table 2 (Continued)

ID	Threats, unwanted incidents
25	Mobile client can be moved between the secure and insecure zone (this assumes use of wireless network, WiFi). The client can be used as a mechanism for communication (transfer of data) between secure and insecure zone. This also includes transfer of malicious software which can affect the computer in the secure zone, without affecting the mobile device (the mobile device then acts as a host for the "parasite")
26	Synchronisation (equivalent of bridging between secure and insecure zone?). During synchronisation, the device can at the same time be connected to the local secure network via the PC, and GPRS or other networks The mobile device can have been infected with malicious software from Internet—malware which is now being transferred to the local network during synchronization (as for 25 above)
27	A backup of telephone memory to computer will also copy the common key from the phone to the backup area (the key will be encrypted with the user's password). Similar to accessing the key on a computer
28	Service is unavailable for the user caused by fall-out of the radio network (GSM/GPRS, etc.), e.g. because of network overload, failures of central infrastructure, etc.
29	Service is unavailable for the user because the phone is outside the service area of the radio network (GSM/GPRS, etc.)
30	Service is unavailable for the user because there may be zones in a hospital where use of some radio networks are not permitted
31	Spam—mass sending off instant messages. Unwanted messages take the attention away from important messages

Two of the threats (10 and 22) were not used in the analysis because we were not able to set a probability level for these.

5.1. Risks related to the use of mobile devices

Many of the most serious threats we identified were related to the possible revealing of confidential information. In addition to all patient related information which by law is confidential, encryption keys used in the communication must also be kept secret (threat 20). If the private key of a user has been compromised, a new key pair must be generated for that user and the new public key must be distributed to all possible communication partners.

Revealing of confidential information can happen in different ways:

- One of the general threats when dealing with mobile units is the possibility that the users lose their devices (threats 1–5 and 9). With a high number of users, it is a large probability that some of them might lose their unit. If a lost device is found it may give unauthorised persons a possibility to try to access the service, and to use the service by impersonating an authorised user. If confidential information exists unencrypted in the memory of the unit, it may be read by unauthorised persons (threat 18). There is also a small possibility that the finder will crack the login passwords for the use of the IM service (threat 4). If a device is lost, there is a difference between the cases when the device is switched on or not.
 - If the device is switched off it has a certain degree of protection by mandatory use of PIN at switch-on (threat 2), but still a possibility that the finder will try to crack the PIN (threat 3). Without the PIN, the finder can replace the SIM (subscriber identity module) of the mobile unit with his own SIM, for which he knows the PIN, and all information which is stored unencrypted in the unit's internal memory will be available to him (threat 18).
 - To handle the switched on case (threats 1 and 9), it is possible to use a time-out mechanism that requires the user to repeat the PIN after the device has been inactive for a certain time frame. The length of the time-out will be a trade-off between security and usability. Con-

stantly repeating the PIN-code will lead to frustration for the users.

- The external memory card of the mobile device may also contain confidential information; it is not protected by the telephone's PIN and it can easily be stolen or lost without the complete device being stolen (threat 5).

The obvious way to reduce these risks is to avoid storing confidential information. The messages of IM are of transient nature, similar to phone conversations. Therefore, it has not been intended to let the system automatically store messages containing patient information. If something really needs to be stored, it must be encrypted.

In addition to the possible revealing of confidential information, the service will be unavailable to the user when losing his mobile device (threats 2 and 6). To avoid long-term unavailability of the service, there should be routines for replacements of lost units.

The service also becomes unavailable if the mobile device loses its connection to the network (threats 28–30). This may happen if there are errors in the network, if the device is outside the service area of the network, or if the device is in an area of a hospital where use of wireless networks is not permitted. Such problems are largely unavoidable.

A more serious threat is related to the possible loss of information from volatile memory as a result of power loss, failures on the device or programming errors (threat 19). The consequence is most severe if the memory contains messages that have not been read yet. But, again, the messages of IM are of transient nature, like telephone conversations, and it is a question whether messages at all should be stored permanently.

5.2. Risks related to malicious software

Another group of serious threats is related to the direct connection between the mobile device and the local secure network, and the corresponding possibility of infection by malicious software:

- The same mobile device can be used both externally via GPRS, and internally inside the secure zone via a wireless network (WLAN) (threat 25). The device is thus a mechanism for communication (transfer of data) between secure and insecure zones. Also malicious software can be transferred in this way. This can be “malware” for PCs in the network, without doing any harm to the mobile device which simply acts as a host for a parasite. Relatively few viruses exist today that can infect mobile devices, but the number is increasing [22].
- The synchronization of the mobile device with the user’s PC is another way of spreading malicious software (threat 26). If the device is using GPRS when it is connected to the PC in the local network, the device may accidentally create a bridging between the two networks. This will open up to attacks from the external network, and allow unfiltered traffic between the networks.
- An intruder can connect to the mobile device via Bluetooth or other network services (infrared, or even SMS/MMS) (threat 8). This can at worst give full access to the mobile device. Attacks can be imposed by memory overflow or exploitation of protocol weaknesses, or running commands that can modify the set-up of the device.

To decrease the risk of infection, limitations should be placed on the use of other services, networks and synchronization. Only the intended type of communication, i.e. the connection to the XMPP server, should be allowed on the mobile device. Bluetooth, and maybe also SMS and MMS, should be deactivated. In addition, we can restrict the users’ access to web and e-mail from the mobile devices, and place restrictions on the types of data that are possible to synchronize. Especially e-mail synchronisation should be restricted. This will reduce the risk of infection, but it will also reduce the functionality of the mobile device. Use of anti-virus software both on the mobile unit and the PC will also reduce the risk. Finally, the mobile device should have a configuration that makes it impossible to use both networks (external and internal) at the same time.

5.3. Risks related to the system architecture

In addition to attacks on the mobile devices, attacks can be imposed on the desktop client, or on the central modules of the service (the proxy, the XMPP-server, and the user directory) and on the (wireless) local network of the secure zone. This may cause different types of problems. The service may become unavailable, messages are not transmitted, or they are transmitted at the wrong time. The presence status and subscription information of the users may also be compromised. This could be serious, e.g. if a user communicates with a psychologist, and the attacker changes the subscription information so that the user is not communicating with his psychologist, as he believes, but with an unauthorised person.

One possible threat to the system is that the proxy is compromised and manipulated by attackers. This can also happen to the server, but that is less likely because the server is placed in a more secure environment. An attacker with control over the proxy could stop, delay, or replay messages (threat 14). These types of attacks are mainly annoyances for the users;

they lose trust in the service and they would use other communication media instead. To overcome this, the server could give a warning to the recipient if a message has a long delay. The recipient of a message could send an acknowledgement for every message received. In this way the sender will be notified if a message is lost. It might be possible for a compromised server to send fake acknowledgments, but by use of PKI (public key infrastructure), the acknowledgments could be digitally signed to make it possible to detect this.

An attacker with control over the proxy could also change the sender or recipient of a message or pretend to be the logged-in user (threats 12 and 15–17). These are more serious threats. To overcome this, the sender of a message can attach a digital signature to the message, making it possible for the recipient to verify that the sender is who he claims to be. By use of PKI-based encryption, only the intended recipient is able to decrypt and read the message. Basically, it should only be possible to address a message to a recipient who is registered as a legal user within this service.

6. Discussion

There are basically four different approaches to handle a risk [23]:

- Accept the risk, in accordance with the organisation’s security policy. These are the risks that are low enough to be acceptable. It is worth remembering that accepting the risk does not mean accepting the unwanted incident indicated by the threat.
- Reduce the risk to an acceptable level. Since the risk is a product of likelihood and consequence, this means to reduce the likelihood, the consequence, or both. It is most often difficult to reduce the consequence of a threat, so the focus should first of all be on reduction of the likelihood.
- Avoid the risk, i.e. not be exposed to the risk, not do the things that could lead to the risk.
- Transfer the risk to a third party (e.g. an insurance company).

Our risk analysis showed that most of the proposed risk treatment measures must be put into function to obtain a service risk level which is acceptable to healthcare. This could, for instance, mean blocking much of the additional functionality of the smartphone, like e-mail and web access, leaving only telephone conversation and our instant messaging service as accessible functionality. Is this a useful solution? Or will the smartphone then be just another gadget in the pocket, in addition to a PDA, a pager, and an ordinary cellular phone?

With our application neither the pager nor an ordinary cellular phone will be needed. The telephone conversation functionality will still be available on the smartphone, and the IM service will compensate for the pager and partly the work related e-mail, and the calendar function will be present on the smartphone. On the other hand, many doctors already use PDA for access to medical reference sources via web, and also to reach patient information (e.g. electronic health record, EHR) while away from their office. They might not accept to have a smartphone where this functionality is blocked.

To fully conclude on the usefulness of the IM service, it must be evaluated in a trial study of the human-computer interaction of this type of application.

Having an electronic service like our MedIMob it seems obvious to automatically store the communicated messages into the patient's electronic health record. One reason that we did not propose this for our prototype is that we compare the IM communication with telephone conversation and thus proposing the documentation to be at the same level. We think that the nature of the messaging, the number of messages and the content of the messages do not make them suitable for automatically storing into the health record system. Not all messages will be related to a specific patient and his condition. However, this should be investigated further in future work, for instance by finding ways to extract and store messages of a certain type or with a certain content.

PKI is the solution to many of the identified threats. Our prototype application has implemented end-to-end encryption at application level, as proposed in the XMPP protocol extension [21], using hybrid encryption with RSA, AES, and SHA-1 as vital algorithms. This has had no noticeable influence on the performance (message throughput) of the service. PKI used with mobile devices imposes some additional challenges. If the private key is stored on the mobile device, it may be revealed in case of loss of the device. An additional protection could be to encrypt the private key with a passphrase. This implies that the passphrase has to be given whenever a message shall be signed or decrypted. This is probably too troublesome. Another approach is to decrypt the private key at log-in to the service, giving the passphrase only once. This means that the key is decrypted whenever the device is in use, and our service is foreseen to be in continuous use and messages can be communicated at any time. The most important precaution will be to have good routines for quick reporting of lost devices and the corresponding revocation of keys.

The use of PIN for authentication is often regarded a weak point, as was also seen in our risk analysis. There is an ongoing development in replacing PIN-based authentication with different biometry methods (finger print, eye scan, etc.). It will be natural to switch over to using these methods when they become more common on mobile devices.

Some alternatives can be seen to the preliminary architecture proposed by Fig. 1. If the IM system is made for one specific hospital only, with the possibility for externals to communicate with this hospital, the use of the national healthcare network is not necessary. In this case the XMPP server can be placed in the internal zone of the hospital's network and the XMPP proxy in the DMZ of the hospital's network. In this case also a VPN solution (virtual private network) between the external mobile client and the internal XMPP server would be a natural solution, thus avoiding the use of the XMPP proxy. On the other hand, if we foresee an IM solution to be used in general within the healthcare sector, the architecture proposed in Fig. 1, with an XMPP server in the national healthcare network, is the most obvious choice.

7. Conclusion

The success of IM in other fields makes us believe that there is a potential for use also in the healthcare sector. Since the

security level supported by existing public instant messaging services are insufficient, we have described a secure architecture for use within healthcare, based on the XMPP protocol. For our service we included the use of mobile devices, because healthcare workers are operating in a mobile environment with rapid changes in their availability status.

During the design phase we carried out a risk analysis of the information security aspects of the proposed architecture and the intended environment. Our goal was to identify security threats to the use of our instant messaging service within healthcare, and suggest acceptable solutions to the threats.

Many of the identified threats are general when dealing with mobile devices and sensitive data; others are threats which are more specific to our application and architecture. We found the most serious threats to be related to the relocation of the mobile device between the open Internet and the secure internal network, and to the synchronisation of the mobile unit with stationary equipment that is connected to the secure network. In these cases there is a risk that the mobile unit can act as a host for attacks against systems in the secure network, and a risk for unintended revealing of confidential information.

The risk analysis showed that to achieve the necessary security level, most of the proposed treatment measures should be put into function. This means that much of the additional functionality of the smartphone needs to be blocked, like e-mail and web access. In addition to ordinary telephone conversation, our instant messaging service should be the only accessible functionality. However, future research should include an assessment of the importance of IM services compared with functionality that may have to be restricted.

A large number of threats to a system are related to the persons' use of the system, not to the technology itself. Technical solutions to security issues will have little effect without the awareness among users [22]. It is therefore important to also focus on non-technical measures for risk reduction. A primary measure is to educate the users, making them aware of the risks and explain the reasons for restrictions imposed by technical solutions and routines for use.

8. Future work

The use of instant messaging and presence services within organisations has been the focus of several evaluation studies [5-8]. To our knowledge, no similar evaluation studies have been performed within healthcare settings with the purpose to understand the usefulness and limitations of IM technology in the healthcare domain.

To conclude on the usefulness of a service like this, a thorough observation and evaluation of the use of a messaging service will be conducted, focusing on computer-supported cooperative work and the human-computer interaction. A preliminary experiment will be performed by use of existing technology: a two-way pager system with the possibility to predefine standard messages. This is additional functionality that can easily be included in the pager and alarm system in use in the hospital department today. The purpose of this observation and evaluation is to obtain an understanding of

the communication pattern and identify adequate message types and the need for presence information.

The results from this study would then be to incorporate the findings into an improved design and development of our MediMob system. An improved design of the system should also focus on offering a secure use of the built-in functionalities of the smartphone (e.g. web access, Bluetooth, and MMS). The possibility for automatically storing (parts of) the messages into the patient's health record should also be investigated.

Summary points

What was known before

- An increasing adoption of instant messaging (IM) as media for personal communication, mainly in private settings, but also in work environments.
- Healthcare has stronger security requirements than many other work environments have.
- Public IM services are not secure enough for healthcare; many of them were never intended for secure communication [2].
- Standards and several methods for risk analysis of information security exist, although few studies report on experience from use of risk analysis in healthcare environment.

What our study added

- It is possible to obtain sufficient security for use of instant messaging (IM) within healthcare.
- Our architecture proposal for an enterprise IM system is an example of this. It is secure enough to be used in healthcare environments, and it supports both mobile and desktop IM clients.
- Risk analysis of our proposed architecture revealed a number of general security threats for mobile solutions, in addition to threats more specific to our architecture.
- Essential functionality of smartphones (e.g. web access, e-mail) should be disabled to ensure sufficient security when dealing with sensitive information on the mobile device.

Acknowledgement

We would like to thank our colleagues at Norwegian Centre for Telemedicine who have given valuable feedback on this paper.

REFERENCES

- [1] J. Stone, S. Merrion, Instant Messaging or Instant Headache? ACM Queue, ACM, 2004, <http://www.acmqueue.com/modules.php?name=Content&pa=showpage&pid=142> (last visited: 27 October 2005).

- [2] F. Langa, More Instant-Messaging Security Holes, InformationWeek, TechWeb, 2001, <http://www.informationweek.com/story/IWK20010927S0021> (last visited: 27 October 2005).
- [3] D. Jacobson, M. Glowacki, Hidden Threats to HIPAA, Palisade Systems Inc., 2003, <http://www.palisadesys.com/news&events/HIPAAstudy.pdf> (last visited: 27 October 2005).
- [4] Top Five Security Risks for Instant Messaging in 2005. IMlogic, 2005, http://www.imlogic.com/pdf/Top5_IM.Security2005.pdf (last visited: 11 October 2005).
- [5] B.A. Nardi, S. Whittaker, E. Bradner, Interaction and outeraction: instant messaging in action, in: CSCW 2000 (Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work), Philadelphia, Pennsylvania, USA, 2000.
- [6] E. Isaacs, A. Walendowski, S. Whittaker, D.J. Schiano, C. Kamm, The character, functions and styles of instant messaging in the workplace, in: CSCW 2002 (Proceedings of the 2002 ACM Conference on Computer Supported Cooperative Work), New Orleans, Louisiana, USA, 2002.
- [7] H.D. Vos, H.T. Hofte, H.D. Poot, IM@[Work] adoption of instant messaging in a knowledge worker organisation, in: Hawaii International Conference on System Sciences, Big Island, Hawaii, 2004.
- [8] H.T. Hofte, I. Mulder, H.D. Poot, D. Langley, I M mobile, where R U? in: ECSCW 2003 (European Conference on Computer Supported Cooperative Work), Helsinki, Finland, 2003.
- [9] M. Bruun-Rasmussen, K. Bernstein, C. Chronaki, Collaboration—a new IT-service in the next generation of regional health care networks Int. J. Med. Inf. 70 (2003) 205–214.
- [10] B.D. Beardmore, Process-Driven, Wireless Computing & Enterprise Messaging in Healthcare, UnBound Technologies Inc.
- [11] J.E. Bardram, C. Bossen, Moving to get ahead: local mobility and collaborative work, in: ECSCW 2003 (European Conference on Computer Supported Cooperative Work), Helsinki, Finland, 2003.
- [12] Directive 95/46/EC on the Protection of Individuals with Regard to the Processing of Personal Data and on the Free Movement of Such Data, European Parliament and Council of the European Union, 1995, <http://www.europa.eu.int/eur-lex/lex/LexUriServ/LexUriServ.do?uri=CELEX:31995L0046:EN:HTML> (last visited: 27 October 2005).
- [13] Health Insurance Portability and Accountability Act, United States Department of Health & Human Services (HHS), 1996, <http://www.hhs.gov/ocr/hipaa/> (last visited: 27 October 2005).
- [14] ISO/IEC 17799: Information Technology – Security Techniques – Code of Practice for Information Security Management, 2nd ed., 2005-06-15, ISO/IEC 17799: 2005(E).
- [15] Y. Stamatiou, E. Skipenes, E. Henriksen, N. Stathiakis, A. Sikianakis, E. Charalambous, N. Antonakis, K. Stølen, F.D. Braber, M.S. Lund, K. Papadaki, G. Valvis, The CORAS approach for model-based risk management applied to a telemedicine service, in: MIE 2003 (Medical Informatics Europe), St. Malo, France, 2003.
- [16] The CORAS Project, 2004, <http://www.coras.sourceforge.net/> (last visited: 27 October 2005).
- [17] Risk Management, Standards Association of Australia, AS/NZS 4360:1999.
- [18] P. Saint-Andre, Extensible Messaging and Presence Protocol (XMPP): Core, IETF XMPP Working Group, 2004, <http://www.xmpp.org/specs/rfc3920.txt> (last visited: 27 October 2005).

- [19] P. Saint-Andre, Extensible Messaging and Presence Protocol (XMPP): Instant Messaging and Presence, IETF XMPP Working Group, 2004, <http://www.xmpp.org/specs/rfc3921.txt> (last visited: 27 October 2005).
- [20] P.E. Kummervold, PasientLink—Project Information, NST, 2004, <http://www.telemmed.no/index.php?language=en&cat=7457> (last visited: 27 October 2005).
- [21] P. Saint-Andre, End-to-End Signing and Object Encryption for the Extensible Messaging and Presence Protocol (XMPP), IETF XMPP Working Group, 2004, <http://www.xmpp.org/specs/rfc3923.txt> (last visited: 27 October 2005).
- [22] A. Shevchenko, An Overview of Mobile Device Security, Viruslist.com, Kaspersky Lab, 2005, <http://www.viruslist.com/en/analysis?pubid=170773606> (last visited: 24 October 2005).
- [23] E. Cavalli, A. Mattasoglio, F. Pincioli, P. Spaggiari, Information security concepts and practices: the case of a provincial multi-speciality hospital, *Int. J. Med. Inf.* 73 (2004) 297–303.