Professional PowerPoint Presentations can Compromise Data Security

Mads R. Dahl [1], Christian B. Høyer [2]
University of Aarhus, Faculty of Health Sciences, Vennelyst Boulevard 6, 8000 Aarhus C, Denmark
[1]Section for Health Informatics, e-mail md@hi.au.dk. [2]Unit for Medical Education

Abstract: Background: The publication of documents on the internet poses a security threat to individuals, researchers, companies, and organizations. The aim of this paper is to describe the extent of this lapse in security, how data can be exposed, and simple measures to keep data safe.
Standard format documents can include Object Linking and Embedding (OLE-objects) which may include hidden data, in some cases sensitive data, such as data from original research or health care information.
Objective: In order to secure confidential data and minimize the risk of publishing data unintended, we give recommendations to both authors of documents and developers of software.
Methods: Google was used to search for Microsoft PowerPoint files. Four increasingly specific searches were performed: 1) PowerPoint files in general, followed by addition of the search terms 2) 'HIV', 3) 'WHO', and 4) 'survey'. The top 250 files returned with each search (totaling 1,000 files) were downloaded and reviewed. Checkpoints were the number of files containing graphs/tables 1) in total, 2) that could not be manipulated, 3) that could be manipulated, and 4) that were based on embedded data (OLE-objects).
Results: With increasing specificity in the search string, the rate of OLE-objects in the presentations showed statistically significantly increases from 30.8% (PowerPoint files in general), to over 62.4% (adding 'HIV'), 57.6% (adding 'WHO'), and, finally, 72.0% (adding 'survey' to the search string). The rate of OLE-objects in the PowerPoint documents in the four groups remained relatively constant, at approximately 1/3 of the files that contained charts or tables.
Conclusions: Uploading PowerPoint files to the internet poses a severe threat to data confidentiality if OLE-objects are used. This may have serious consequences for patients and researchers, among others. We recommend the software industry to take action and show responsibility for a more secure interoperation between applications.
Keywords: Microsoft Office, PowerPoint, Sigmaplot, data security, privacy, information systems, informatics.

1. Introduction

Microsoft PowerPoint dominates the market for presentation software as 95% of all digital presentations are made using the PowerPoint software [1]. PowerPoint is sold as part of most versions of the Microsoft Office suites, which are installed on hundreds of millions of computers: More than 120 millions of licenses to the Microsoft Office 2007 suite *alone* have been sold, according to Microsoft [2]. The widespread use of PowerPoint, as well as the rest of the Microsoft Office suite, has obvious advantages, especially the great portability of documents. Combined with the ease of distribution of files via the internet this contributes greatly to potential knowledge sharing among commercial enterprises, public authorities, associations, educational institutions, and health care organizations [3]. Apart from the simple exchange of information between friends and colleagues, it is also possible to explore the internet for sources of information not otherwise found due to factors such as geographical location [4]. Thus, a literature search on the internet for scientific papers, reports, and presentations has become a standard method of obtaining knowledge and inspiration [3] [5].
However, the benefits of easy information sharing are accompanied by downsides: confidentiality and security can be compromised [3]. Security measures like encryption, password protection, and access control can be implemented to protect data from unauthorized access [6] [7], but do not hinder security breaches due to human error (a classic example is data e-mailed to the wrong recipients [8]). In the following text, the word *hidden data* is used to refer to all kinds of data that are not intended by their author to be publically available.
Usually, most files can be edited and re-distributed without any restrictions. In addition to their subject content, they often include information about the author, company, and changes made in the process of making the document (metadata). Furthermore, interoperability between software applications often means that data from one application (e.g., a spreadsheet) can be found in another (e.g., a presentation). In other words, if the data stored in the spreadsheet include, for example, detailed information about health, religion, social security number, financial reports, and collaborators, then these data may be available in the presentation. If the presentation is uploaded to the internet, all these data will be freely available to internet users worldwide. Confidentiality, Integrity and Availability are three key aspects in information security [9]. Documents uploaded to the internet have a maximum availability and in if the document has embedded confidential data it takes the confidentiality to a minimum. The integrity of the data is often reliable and correct, since the author of the presentation document is also the author of the data or has unlimited access to the original data [10] [11].
The purpose of this paper is to describe the degree that hidden data embedded in PowerPoint files are found on the internet, to give examples of the types of data found, and to

Received December 15, 2008

make recommendations about how to prevent the risk of inadvertently making hidden data publically available.
Certain technologies that are individually very useful can potentially be harmful if combined. Examples include 1) the use of Object Linking and Embedding (OLE), 2) the use of the Microsoft Office suites, and 3) the easy dissemination and retrieval of files on the internet.

Object Linking and Embedding
The OLE technology integrates functions from different applications and allows for interoperation among software applications. For example, visual presentations of data from a spreadsheet can be used directly in a document in a word processor, which could not normally produce these illustrations itself.
OLE works in two ways. Data can be embedded in the document and thereby distributed with the document. Alternatively, a link can be established between the original source and the target document that enables data to be updated between the documents. In this case, if data are updated in the original document, data are updated as well in the target document. Often, data are embedded and linked simultaneously. The OLE system has been increasingly relied on to strengthen co-operation between computer applications. It has, for example, been set as the default setting when using the copy/paste function to export charts from Excel 2003 into PowerPoint 2003 or Word 2003.

Microsoft Office and the internet
The use of PowerPoint, together with the other applications in the Office suites, goes beyond what most people can probably understand. In 2004, Simons estimated that more than 20,000,000 PowerPoint presentations were given every day. Two years later, the estimated number had increased to 30,000,000 [12]. More than 4,000,000 PowerPoint presentations were available in 2006 on the internet [12]. Today (May 2009), the number of PowerPoint files found on Google exceeds 8,740,000. Likewise, the numbers of Excel and Word files exceed 12,300,000 and 60,200,000, respectively. The number of PowerPoint files indexed by Yahoo exceeds 24,000,000, and the numbers of files are increasing each day.

Searching the internet
The use of modern search engines, such as Google, MSN, or Yahoo, makes it easy to find specific file types by searching for their extensions: .doc, .xls, and .ppt for Word, Excel, and PowerPoint, respectively. The use of specific terms, like *clinical trial*, *religion*, or *social security number*, will target the search and may reveal hidden, potentially sensitive data.

Uncovering data
The danger is when the author of the file inadvertently embeds confidential data in the documents and thereby exposes these data to users of the internet. The author uses the simple copy-paste function (Ctrl-C followed by Ctrl-V), activates the OLE technology, and uploads the entire dataset to the internet.
The recognition of OLE-objects in files is easy. Often graphs or charts have their default layout from Excel, and are therefore quickly identified as possible OLE-objects. When double clicking on a graph, the spreadsheet opens (Figure 1).

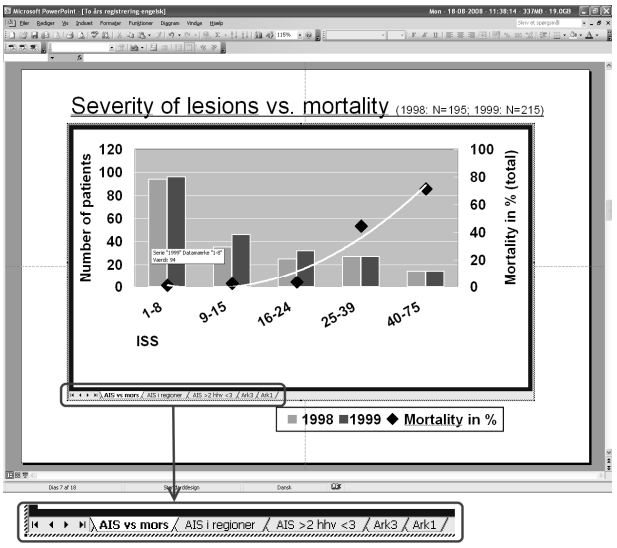

Figure 1. Screen-dump showing an embedded Microsoft Excel document in a Microsoft PowerPoint presentation. When double clicking on the graph, the spreadsheet is revealed. This can be seen by the appearance of the row of tabs below the figure (enlarged).

The data used in constructing the graphs can now be found when clicking on the different tabs (Figure 2).

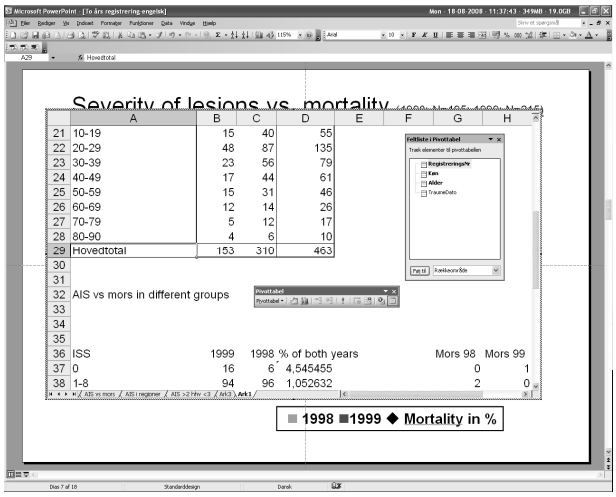

Figure 2. Clicking on the tab 'Ark 1' (Danish for Sheet 1) reveals all the data used for construction of the graph. The data shown originated from a study performed by an author (CBH) of this paper.

2 Methods

Only PowerPoint presentations were considered in this study. Google was used as the search engine due to its widespread use. All files were identified and downloaded on August 28th 2008 and the three cases described in this paper were re-downloaded 22nd January 2009. On these two dates, Google estimated that 5,610,000 and 7,730,000 PowerPoint

Received December 15, 2008

files were available on the internet respectively. The sample size was chosen to be 1,000 unique files, that is, 250 files for each of the four searches. Duplicate files were included only in the first search in which they appeared. In cases of duplicate files, or defective files, the next file in the search results was included to reach a total of 1,000 unique files.

The check points used for evaluation were:

1) The number of files containing graphs or tables,

2) The number of files with graphs or tables that could not be manipulated (such as image files),

3) The number of files with data that could be manipulated (such as graphs made by entering numbers in tables directly in PowerPoint), and

4) The number of files with embedded data documents (OLE-objects).

Search strategy

Searching for specific file types in Google is done using the "filetype:" command. In this case, when searching for PowerPoint files, the command was "filetype:ppt." The four searches made were 1) any PowerPoint file, 2) files found when adding 'HIV' to the search, 3) files found when adding 'WHO', and finally, 4) files found when adding 'survey'.

Hidden data

It was decided *not* to estimate whether the data embedded in the files was sensitive or not. The reason is that data that may seem innocuous to non-experts may actually be very sensitive to experts in the exact field from which the data originate. However, it was decided to contact the authors of the examples chosen, to clarify whether the published data were in fact meant to be publically available or not.

Statistics

The comparison of the observed values between the groups was done using χ^2; the level of significance was chosen to be $\alpha = 0.05$.

3.Experiment Results

A total of 1,006 .ppt files were downloaded using the Google search engine to obtain a total of 1,000 unique files, as three files appeared in more than one of the searches and three other files were defective. For each search, Google estimated the number of .ppt files available (5,610,000, 61,200, 31,500, and 8,770, respectively).

Tables 1 & 2 describe the distribution of files, including the presence of graphs or tables and whether these could be manipulated or were OLE-objects. The proportion of tables or graphs in the PowerPoint files showed a statistically significantly increase with increasingly specific search strategies (χ^2, $p < 0.05$). The proportion of OLE-objects in those files with tables or graphs remained relatively constant, at approximately 1/3 of the presentations (29.5 - 39.0%) ($\chi 2$, $p = 29.9$).

Search string	filetype: ppt (n=250)	HIV filetype: ppt (n=250)	WHO HIV filetype: ppt (n=250)	WHO HIV survey filetype: ppt (n=250)
- graphs or tables	173 (69.2%)	94 (37.6%)	106 (42.4%)	70 (28.0%)
+ graphs or tables	77 (30.8%)*	156 (62.4%)*	144 (57.6%)*	180 (72.0%)*

Table 1. The distribution of 1,000 files with and without graphs or tables. * Statistically significant difference, $p < 0.05$. There were found graphs or tables in 557 of the 1,000 unique files downloaded (55.7 %).

Search string of + graphs or tables	filetype: ppt (n=77)	HIV filetype: ppt (n=156)	WHO HIV filetype: ppt (n=144)	WHO HIV survey filetype: ppt (n=180)
No metadata	24 (31.2%)	53 (34.0%)	56 (39.9%)	50 (27.8%)
Adjustable data	23 (29.9%)	57 (36.5%)	44 (30.6%)	71 (39.4%)
Embedded data (OLE)	30 (39.0%)	46 (29.5%)	44 (30.6%)	59 (32.8%)

Table 2. Graphs or tables without metadata could be pictures (.jpg, .tiff, and .bmp) pasted into the presentation, while graphs or tables with adjustable data are constructed by using the innate PowerPoint function to make a graph by entering data directly in a table in the file. Embedded data (OLE-objects) are tables or graphs containing data hidden behind the object produced in another program.

Received December 15, 2008

Three examples of potentially sensitive data found in the PowerPoint files are described in the following Sections.

Case 1: Business guide to partnering with NGOs and the United Nations.
This slideshow included information about 309 companies in 100 countries, regarding their partnerships with 671 non-governmental organizations (NGOs) [13].
The embedded data included detailed contact information for 292 persons (name, telephone numbers, and e-mail address) and a listing of the 987 relations between the companies and the NGOs. Also included was a listing of each company's priorities when engaging in partnerships with NGOs, as well as companies who did not engage in partnerships and their reasons for this. According to the author (personal communication), the embedded data were not for general distribution. The presentation could be found and downloaded on January 22nd 2009.

Case 2: The cost of HIV in Africa
Another slideshow included detailed figures about the expenditures from 'The Global Fund to Fight AIDS, Tuberculosis and Malaria' in 2004 [14]. Specifications about applications for funding of a total of US $1,011,888,777 were available, as were the specifications about the grants that were approved (a total of US $346,201,782). In the embedded worksheets, it was possible to find information about grants approved to organizations such as UNICEF, WFP, WHO, and USAID. The presentation was not made available on the internet by the author, but by some other source unknown to the author. According to the author (personal communication), the data embedded in the file are now outdated. The presentation could be found and downloaded on January 22nd 2009.

Case 3: Stigma and discrimination associated with HIV and AIDS
This presentation was published November 13th 2006 in connection with a conference at the University of Amsterdam, Holland [15]. The presentation summarizes a survey study of HIV related discrimination from six countries in Asia. On five out of thirty-two slides charts depicting specific results of the survey can be found. Upon analysis of the individual charts they were found to be imported from the program Sigmaplot 8.0 (Systat Software, US) having embedded data apart from what was apparent in the charts themselves. The data embedded in the charts did in no way compromise the integrity of the participants in the survey. The presentation could be found and downloaded on January 22nd 2009.

Discussions
We analyzed 1,000 PowerPoint files as data for this paper. The files were selected by searching with Google using four search strings, and subsequently downloading the first 250 unique files in each search.
Tables or graphs were included in 55.7% (557/1,000) of the examined files. The OLE technology was used in 32.1% (179/557) of the files with graphs or tables. The two first cases described above are violations of good data management practices. Case 3 illustrates that embedded data charts is not isolated to Microsoft products, but a more general problem including several other popular software products.

Consequences
The positive effects of knowledge sharing in the 21st century cannot be disputed. However, in some cases the sharing of data can be disastrous to the presenter's intellectual property. Researchers may destroy their research project by exposing confidential data, such as the GPS-coordinates of an archaeological site. Others may forfeit the right to take out patents or publicize findings, as the results are already available to the public. Individuals' private or work relations may be compromised if information about their sexual behavior, religious affiliation, economic status, or other personal information is revealed. One example would be if a member of a church with strong religious sentiments against abortion was revealed to have had an abortion; another example would be if a clinical doctor presented demographics of his HIV-positive patients based on an embedded table that included the individuals' names and social security numbers. Security risks may arise to organizations such as governmental or non-governmental organizations if detailed information about their activities is publically available. As example, the security of a humanitarian aid program could be severely compromised by information about which parties are being supported in an armed conflict or routes for evacuation in case of emergencies. Sensitive information about commercial companies' future activities, if revealed to competitors, could jeopardize their position in the market.
This list of possible consequences may seem far-fetched, but real examples exist. One is when an inadvertent disclosure about finances and upcoming services for Google Inc. resulted in an almost 2 percent drop in Google's shares in the following hours [16]. Another example is how security concerns from the American Federal Bureau of Investigation were allegedly leaked via an internal PowerPoint presentation [17]. In this way, all other data management procedures implemented over the years, such as high level encryption software, firewalls, and server protection measures, may be meaningless as information can be found in documents published on purpose [18] [19].
Other researchers have described the dangers of exposing data in PowerPoint presentations: examples include insufficient masking of patient data when using radiological images, inexpedient naming of images used in the presentation, or patient data in speaker notes [20] [21]. Weadock and colleagues scrutinized 200 PowerPoint files found by Google and found 82 images (41%) containing patient data. Of these, 31 cases included the name of the patient [20]. Much to our surprise all cases described in this paper could still be found and downloaded via Google four months after we informed the responsible author.
Software developers have made interoperation between applications possible and very easy. However, the usa of OLE-technology as default setting in PowerPoint when inserting charts may compromise data security on behalf of assumed usability.

Limitations
The aim of this study was to quantify to which extent it was possible to locate data embedded in PowerPoint

Received December 15, 2008

presentations. While the possibility to publish hidden, potentially sensitive data by uploading PowerPoint presentations on the internet is shown, we did not include other standard document formats such as Word or Excel files. Consequently, it is not possible to show whether the same danger exists when these types of documents are published. However, as the technology used (OLE-objects) is the same for these documents as in PowerPoint documents, we believe it is reasonable to warn about publishing these document files as well. Furthermore, our analysis did not include whether metadata (for example, information about authors, institutions, office, reviewers, or authorization of the document) or speaker notes in the presentations included hidden data. It can therefore be argued that our results underestimate the magnitude of the problem. If this is the case, it is even more important to focus on the danger of publicizing Office documents on the internet. As we did not make a specific assessment of the possible sensitivity of all OLE-objects, it can be argued there is no supporting evidence regarding the threat to data confidentiality. However, given the examples cited, we believe that our conclusions are supported.

Presentations and documents can be found on the internet using a search engine such as Google. This strategy has been used by other researchers exploring the same subject [20]. The number of hits on a particular search string that has been restricted to search for only a specific file type reflects the number of files indexed by the search engine and not the true number of files. The majority of presentations and documents are “invisible” to the search engines, i.e., the web pages having limited access for search engine robots or the documents are placed in subfolders. Thus, the total number of files available on the web is difficult to estimate.

Recommendations

On the internet it is possible to find apparent “solutions” to these problems. Microsoft has published an “add-in” to Microsoft Office 2003/XP [22]. Microsoft describes this add-in as a permanent way to remove hidden data from Microsoft Office documents. However, this is not the case, as the OLE-objects are not altered in any way. Other tools that claim to be able to remove unauthorized private information are available, but do not address the problem with OLE-objects: they only remove metadata, including author, version, and tracked changes [23] [24].

The most basic method to take precautionary measures against unintentionally revealing confidential or sensitive data is to *ensure complete separation between the data and the documents used to present results* [25].

Other easy provisions are:

- Insert figures and tables by using the menu “Edit”, “Paste special”, “Picture”, and thereby avoid pasting figures and tables by the “copy-paste” method (shortcuts Ctrl-C followed by Ctrl-V).
- Avoid the publication of PowerPoint files (and Word, Excel, and Access files, among others).
- Convert presentations (and other documents) to a format that does not include embedded data.
- Avoid the implementation of software solutions claiming to anonymize documents, unless thoroughly tested to meet the necessary requirements.

In conclusion: this study has shown that filetype targeted and topic specific internet searching with subsequent download can be conducted. We found that millions of PowerPoint files can be found on the internet using simple search strings in standard search engines and approximately one out of five of the 1,000 files downloaded in our study contained OLE-objects that in some cases included specific information about persons or finances. We found that interoperation between PowerPoint and charts produced in Excel or Sigmaplot could embed the original datasheets. Thus, OLE-objects may be considered as a major threat to data confidentiality, as the inadvertent use of this technology can bypass all other security measures. We recommend the software industry to take action for a more secure interoperation between applications and acknowledge the security risk of the OLE-objects. The most basic precautionary measure against the publication of confidential and sensitive data, however, remains to be ensuring the complete separation between raw data and documents used to present results.

References

[1] Parker I. "Absolute Powerpoint - Can a software package edit our thoughts?", *The New Yorker*, 2001

[2] Elop, S. "Financial Analyst Meeting 2008", *MSFT Investor Relations*, last update 24-7-2008, accessed 14-8-2008.

[3] Dieng R, Corby O, Giboin A, Re MR. "Methods and Tools for Corporate Knowledge Management", *Int J Human-Computer Studies*, 15. pp. 14-17, 1999.

[4] LaPorte RE, Linkov F, Villasenor T, Sauer F, Gamboa C, Lovalekar M *et al*. "Papyrus to PowerPoint (P 2 P): metamorphosis of scientific communication", *BMJ*, 325. pp. 1478-1481, 2002.

[5] Delamothe T, Smith R. "Moving beyond journals: the future arrives with a crash", *BMJ*, 318. pp. 1637-1639, 1999.

[6] Jennett P, Watanabe M, Igras E, Premkumar K, Hall W. "Telemedicine and security. Confidentiality, integrity, and availability: a Canadian perspective", *Stud Health Technol Inform*, 29. pp. 286-298, 1996.

[7] Maamir A, Fellah A, Salem LA. "Controlling Information Flow in Object Oriented Systems", *Journal of Information Assurance and Security*, 2. pp. 140-146, 2008.

[8] Gillam L, Cooke N. "Intellectual property escaped with the email? Press F1 for help", *Journal of Information Assurance and Security*, 1. pp. 16-26, 2008.

[9] Pharow P, Blobel B. Public key infrastructures for health. *Stud Health Technol Inform*, 96. pp. 111-117, 2003.

[10] Berman JJ, Moore GW, Hutchins GM. "Maintaining patient confidentiality in the public domain Internet Autopsy Database (IAD)", *Proc AMIA Annu Fall Symp*, pp. 328-332, 1996.

[11] Collmann J, Coleman J, Sostrom K, Wright W. "Organizing safety: conditions for successful

Received December 15, 2008

information assurance programs", *Telemed J E Health*, 10. pp. 311-320, 2004.
[12] James KE, Burke LA, Hutchins HM. "Powerful or Pointless? Faculty Versus Student Perceptions of PowerPoint Use in Business Education", *Business Communication Quarterly*, 69. pp. 374-396, 2006.
[13] Skovby, H. "Business guide to partnering with NGOs and the UN", *Internet*, last update 5-7-2007, accessed 22-1-2009.
[14] Alban, A. "The cost of HIV in Africa", *Internet*, last update 11-5-2007, accessed 22-1-2009.
[15] Reidpath, D. D. "Is there a universal measure of the burden of disease?", *Internet*, last update 13-11-2006.
[16] Kopytoff, V. "Google's gaffe reveals internal secrets: Notes inadvertently offer a look at financial plans, future product", *San Francisco Chronicle/SFGate*, last update 8-3-2006, accessed 2-10-2008.
[17] Anonymous. "FBI Fears Chinese Hackers Have Back Door Into US Government & Military", *Abovetop-secret.com*, last update 21-4-2008, accessed 3-10-2008.
[18] Norifusa M. "Internet security: difficulties and solutions", *Int J Med Inform*, 49. pp. 69-74, 1998.
[19] Albisser AM, Albisser JB, Parker L. "Patient confidentiality, data security, and provider liabilities in diabetes management", *Diabetes Technol Ther*, 5. pp. 631-640, 2003.
[20] Weadock WJ, Londy FJ, Ellis JH, Goldman EB. "Do radiology and other health care presentations posted on the Internet contain accessible protected health information?", *Radiology*, 249. pp. 285-293, 2008.
[21] Yam CS. "Removing hidden patient data from digital images in PowerPoint", *AJR Am J Roentgenol*, 185. pp. 1659-1662, 2005.
[22] Microsoft Corporation. "Office 2003/XP Add-in: Remove Hidden Data", *Internet*, last update 7-8-2008.
[23] Smart PC Solutions. "Metadata Analyzer - Analytical tool for checking MS Office documents", *Internet*, last update 16-9-2008.
[24] 3BView - Secure Document Exchange. "3BClean", 2006.
[25] Ilioudis C, Pangalos G. "A framework for an institutional high level security policy for the processing of medical data and their transmission through the Internet", *J Med Internet Res*, 3. pp. E14, 2001.

Author Biographies

Mads Ronald Dahl: PhD in Immunology and Medical Microbiology (Leicester, UK, 2004); Master in Health Informatics (Aalborg University, DK, 2007); M.Sc. in Biotechnology (University of Aarhus, DK, 2000). Member of the Danish computer Society. Currently leader of the Section of Health Informatics, University of Aarhus, Denmark. Previous publications and research in Bioinformatics, Health Informatics, Usability and data security.

Christian Bjerre Høyer: Medical doctor, graduated from Faculty of Health Sciences, Aarhus University, Denmark, 2004. Main field of study: Medical education of junior physicians in resuscitation, crisis resource management, and communication in professional teams in healthcare.

Received December 15, 2008