Computer security clinical trials: Lessons learned from a 4-month pilot study

Fanny Lalonde Lévesque
École Polytechnique de Montréal
fanny.lalonde-levesque@polymtl.ca

José M. Fernandez
École Polytechnique de Montréal
jose.fernandez@polymtl.ca

Abstract

In order for the field of computer security to progress, we need to know the true value of different security technologies and understand how they work in practice. It is important to evaluate their effectiveness as they are being used in an ecologically valid environment. To this end, we postulate that security products could be evaluated by conducting computer security clinical trials. To show the feasibility of such approach, we did a 4-month proof of concept study with 50 users, that aimed to evaluate an anti-malware product. In this paper, we present the study we performed and provide lessons learned and recommendations on the challenges, limitations and considerations of conducting computer security clinical trials.

1 Introduction

Insecure computers cause numerous problems on today's Internet. From computer-mediated blackmail to corporate data breaches to the spam generated by million-machine botnets, we are regularly reminded that our computer systems are not secure enough. To try and address this problem, companies and academic researchers have developed numerous technologies and techniques for defending computer systems. Rather than propose yet another technology for improving computer security, here we are concerned with a more fundamental question: how do we evaluate the value added of security technologies as they are deployed?

Our insight here is that to evaluate the efficacy of computer defences in practice, we need to observe them as they are being used in real conditions. As many infections rely on direct or indirect user action, such as opening an email attachment, visiting a malicious Web site (drive-by download) or not updating the operating system, we suggest that these actions (or inactions) could lead to a vulnerable system weeks or even months after they occurred. Therefore, these situations can not be accurately reflected in lab-based studies. Moreover, users may also impact how security technologies perform. For example, ignoring the dialog boxes or misconfiguration of the product may result in a compromised system.

Given that users are involved in both the infection and the protection process, security products should be tested with actual users that interact with the product in a realistic environment over an extended period of time. By including the user in the evaluation, one can gain better insight of how security technologies are used and how external factors, such as the environment, the system configuration and user behaviour, can influence their effectiveness.

In previous work, we described this approach to the CSET community [23] and later demonstrated it in a proof of concept clinical trial involving 50 subjects over a 4-month period [9, 10, 11]. While this study and its results have been published elsewhere, in this paper we provide previously undisclosed details on the methodology employed, discuss ethical and privacy considerations and provide lessons learned on the conduction of future computer security clinical trials. The rest of this paper is organized as follows. Section 2 presents current testing methods of security technologies. Section 3 describes the concept of computer security clinical trials and the study we conducted. In Section 4, we discuss the lessons learned and recommendations based on our pilot study. We conclude in Section 5 with a discussion on the limitations of this methodology, other potential applications of this kind of study and future work.

2 Security Product Evaluation

Security products are an important line of defence of information system protection against current threats. They can take many forms, such as anti-malware product, anti-spyware product, firewall, intrusion detection system (IDS), etc. Testing how these products are effective at protecting end-users and their system is therefore crucial.

Although there are several methods for testing security technology, they do not reflect the true performance of products in practice. Products are often rated on their features and tested in controlled environment by being exposed to known threats. Typical methods for evaluating anti-malware products are based on scanning collected or synthesized malware along with legitimate programs. One major issue with these tests, known as the sample selection problem [1], is that the sample is often too small, inappropriate, and unvalidated [4, 7]. Firewalls are evaluated by automated penetration testing against tools that exploit security vulnerabilities or by design-oriented testing, that aimed to verify the features provided by the product [16]. These tests, however, are not suitable for firewalls that are highly adaptable and user customizable. Moreover, penetration tests lack to simulate real-life attacker behaviour [12]. Regarding IDS, the evaluation is performed by manual or automated testing against data sets of traffic traces with attacks and intrusions. While artificially generated traces do not contain a sufficient variety of attacks instances and behaviours based on realistic data [13], benchmarking data sets based on real data presents many privacy concerns and legal restrictions [14]. To address these issues, researchers have proposed to use sanitized real traffic [14, 21, 24] or traffic generated from emulated user profiles [3, 18, 19].

While lab-based evaluations can provide insight on specific features, these tests are not representative of real-life situations as they can not account for external factors that may influence the performance of security technologies. Giving the evolving nature of threats and the fact that users play a much more preponderant role in the process of infection, security products should be evaluated following a human-in-the-loop approach. To partially address this issue, Vrabec and Harley [26] suggested to conduct testing of anti-malware products by emulating user-specific testing scenarios. As a first attempt, PC Security Labs conducted an AV test [17] to measure the defence efficiency of AV solutions against 7 different user groups: Internet addict, network businessman, socializer, basic user, gamer, self-presenter and infrequent user. Their test confirmed that AV solutions perform differently depending on the user's profile. While testing with user emulation has just been introduced in the field of anti-malware research, many similar experiments [3, 18, 19] have already been conducted to evaluate IDS systems. Even though this approach can offer a more realistic testing environment, it is impossible to simulate all users' behaviour.

Therefore, real users should be included in the evaluation process of security products if we want to better understand how they perform in real world settings and how users interact with them. For example, Shams *et al.* [22] performed a lab test where anti-spyware products were being used regularly by an end user. Egelman *et al.* [2] added another layer of realism by conducting a lab-based user study to evaluate the effectiveness of Web browser phishing warnings and examine if, why and how they failed users.

We believe, however, we must do more if we are to make significant strides in improving the security of the Internet. As laboratory studies are often challenged with ecological and external validity issues [8], what we need then is a methodology that will allow us to evaluate the interactions between humans, malware, and security products with much greater ecological validity. Much as drugs and other medical interventions are studied first in the lab and then later in the clinic, we have proposed the use of computer security clinical trials in our past work to evaluate security products [23].

With computer security clinical trials, security solutions would be installed and monitored on systems in regular use by regular users. Data would be gathered on the performance of the security product in protecting the system and on how the user interacted with the system during this time period. By correlating user behaviour, application use, system configuration and security software activity, we can gain insights into the interactions between all three in an ecologically valid context.

3 Computer Security Clinical Trials

The 4-month study we conducted [9, 10, 11] involved 50 subjects whose laptops were instrumented to monitor possible infections and gather data on user behaviour. We monitored real-world computer usage through diagnostics and logging tools, monthly interviews and questionnaires, and in-depth investigation of any potential infections. By conducting this first study, we wanted to: 1) develop and test the validity and viability of a new methodology for the evaluation of security products; 2) determine how malware infects computer systems and characterize sources of malware infections; and 3) determine how factors such as the configuration of the system, the environment in which the system is used, and user behavior affect the probability of infection of a system.

3.1 Protocol certification and approval

The first step of the experiment was to seek the approval of the project to the two relevant university instances: the *Comité d'évaluation des risques informatiques* (CERI, i.e. the computer security risks evaluation committee) and the *Comité d'éthique de la recherche* (CER, i.e. research ethics review committee).

The CER is equivalent to other ethics review boards found in universities worldwide. It is mandated by the Canadian government research funding agencies, in our case the National Science and Engineering Research Council (NSERC), that all universities receiving research funds from them have such a committee in order to review all research projects involving human subjects. Similar requirements exist for research conducted with animals, involving risks to the environment, biological hazards, etc. It is typically multi-disciplinary in nature, and composed of peer researchers in areas involving such research, university administrators and external experts on research ethics.

While research involving computer security-related risks, such as that involving malware or discovery of dangerous computer vulnerabilities, is not covered in the Responsible Conduct of Research policy of the funding agencies, our institution, the École Polytechnique de Montréal, decided that it was important to address the issue. Accordingly, a self-imposed policy on the conduct of research that involves or may involve computer security risks was drafted and promulgated in 2009, and a committee was created to evaluate and manage the risks associated with such projects, the CERI [20]. The CERI has the responsibility to proceed with the evaluation of research projects that entail or could potentially entail computer security risks for the institution, its collaborators or any other entity/individual. These risks could include damage to the institution's infrastructure and information assets, financial losses to the institution, harm the institution's reputation or affect the confidentiality, integrity and availability of the institution's data, or that of a third party. The CERI is composed of a computer security expert from industry, the director of university IT services and peer researchers in computer science.

Our project had to be examined and approved by both committees, the CER and the CERI. In order to lighten the administrative burden, and in order to leverage the computer-specific expertise of the CERI, the university research office decided to have the CERI review the project first. By doing so, it gave the CERI the additional mandate, beyond its original one of evaluating computer security risk, of also examining privacy issues concerning subject data and advice the CER accordingly, as the latter did not and does not normally include computer scientists, thus lacking the expertise required to do so.

Computer risks

At first the CERI wanted to impose the duty of due care of users in the case of infections, potentially in the form of 24/7 technical support. However, this requirement was waived when the committee was convinced that the subjects were not going to be significantly more at risk of getting infected that the average home computer user who had installed an up-to-date reputable anti-virus, as it was going to be the case in the study. Nonetheless, this requirement and our ability to provide 24/7 support (due to lack of resources), prevented us from including a control or "placebo" group in the study that would have had no anti-virus installed.

We provided users an AV product that was centrally managed on our own server to guarantee high-availability. The AV software was updated daily and configured to perform a full scan of the computer everyday, in order to provide an equal or better level of protection that average corporate or home users would have. Should the AV detects an infection, it would be automatically neutralized. However, in the event that our diagnostics tools detected an infection on the computer, a procedure was given to the user so he could neutralize the threat by himself.

Giving that the experiment implied manipulation of malware files, special precautions were taken in order to protect the university's IT infrastructure. All malicious or potentially malicious files were first encrypted and copied on DVDs, before being stored in the high security zone of the laboratory. Moreover, all the computers were analysed by being connected to an isolated network in order to prevent any contamination of the university network.

Ethical and privacy considerations

Following the CERI certification, the CER had to approve our recruiting procedures, the experimental protocol, as well as the measures adopted to guarantee user anonymity and confidentiality of the data collected.

To guarantee the anonymity of the participants, we assigned each user a unique number associated with his computer. The only personal information kept for administrative and financial purposes was the participant's name, email address, and telephone number. This information was only accessible by the project leader and was destroyed 3 months after the end of the study.

All raw data and statistics generated during the experiment were anonymized. Since this data was only linked to the user ID, it was not *a priori* possible to identify the subject with the information collected. The data was stored in a locked cabinet in the high-security zone of the laboratory, which is protected with three-factor authentication (biometrics, PIN and ID card). This work zone was completely isolated from the Internet and the university network. The security policy of the laboratory was also applied to the deletion of all personal data related to the experiment. This policy applies to all information whether on paper or electronic media, and conforms with Government of Canada information security standards.

Only authorized personnel within the context of the project was able to access the data. In the event we wanted to share the data with another researcher, he had to agree to comply to the university computer risks and ethics policy. Moreover, the data collection was bound to the purpose of the project's research objectives. Finally, if we had inadvertently discovered information leading a reasonable person to believe that a (serious) crime had been committed or was about to be committed, we would have been required by law to advise the appropriate authorities (law enforcement agencies, etc.). Fortunately, this was not the case in this experiment.

3.2 Equipment

The laptops we provided to the subjects had all an identical configurations, with the following software installed: Windows 7 Home Premium; Trend Micro's OfficeScan; monitoring and diagnostic tools including HijackThis, ProcessExplorer, Autoruns, SpyBHORemover, SpyDLLRemover, tshark, WinPrefetchView, WhatChanged; and custom Perl scripts developed for the purpose of the experiment.

The scripts automated the execution of the tools and compiled statistical data on the system configuration, the environments in which the system was used, and the manner in which it was used. The data compiled by our scripts included:

- The list of applications installed;
- The list of applications for which updates are available;
- The number of Web sites visited per day;
- The number of Web sites visited by categories per month;
- The number and type of files downloaded;
- The list of browser plug-ins installed;
- The number of different hosts to which the laptop communicated;
- The list of the different locations from which the laptop established connection to the Internet;
- The number of hours per day the laptop is connected to the Internet;
- The number of hours per day the laptop is on.

Before deployment, we benchmarked the laptops by running tools and recording the output. The recorded information included: a hash of all files plus information about whether the files were signed; a list of auto-start programs; a list of running processes; a list of registry keys; a list of browser helper objects (BHO); a list of the files loaded during the booting process; and a list of the prefetch files.

Regarding the anti-malware product, it was centrally managed on our server in a manner similar as is usually done for corporate installations to centralize distribution of signature file updates. All the AV clients installed on the laptops were thus sending relevant information to our server on any malware detection or suspected infection as they occurred.

3.3 Experimental protocol

Subject recruiting

We did the recruiting by advertising the experiment on the Université de Montréal campus (which includes the Ecole Polytechnique engineering school and the HEC business school) the using posters and newspapers. Even though the recruiting process was centered on the university campus, the study was open to everyone. Interested participants were invited to visit a designated Web site in order to obtain more details and fill a short on-line questionnaire that we used to collect initial demographic information such as gender, age, status and field of activity. The only exclusion criteria was to be 18 years-old and over. In less than three weeks, we received 131 registrations of potential volunteers.

Given our limitation on population size (number of laptops available), an important issue was to select a sample of 50 users as representative as possible of the general population of Internet users. Due to the over-representation of students and the limited number of candidates, we randomly selected users based on their characteristics. While this approach was suitable for a pilot study, recruiting for larger-scale trials should be more rigorously structured, as is the case for medical clinical trials.

In-person sessions

Users were required to attend 5 in-person sessions: an initial session where they received their laptop and 4 monthly sessions where we collected the data and analyzed the computer. Participants were invited to book their appointments via an on-line calendar system hosted on our server. To encourage subjects to remain in the study, we paid them for each session they attended. If they completed all sessions, a sum equivalent to the purchase cost of the laptop (borne by the users) was reimbursed, along with a small additional compensation.

Initial session The intend of this short session was to obtain the subjects informed consent and provide them their laptop. Each user had to read and sign the informed consent form in order to confirm his participation in the study. Thereafter, the laptop was sold at a reduced price to the subjects. This option was chosen for legal reasons but also to foster user ownership of their computer, in the hope of reducing experiment bias in user behaviour. The only restrictions imposed were that they were not allowed to do the following during the study: i) format the hard drive, ii) install another operating system, iii) delete our tools and the data collected, iv) install another AV product, and v) create a new disk partition.

Participants were also invited to answer an initial questionnaire in order to collect general information of their profile, such as gender, age group, status (worker, student, unemployed), field of activity (computers, applied sciences, pure sciences, art and sciences) and level of computer expertise.

Monthly sessions During the monthly sessions, participants were asked to answer an on-line questionnaire. The aim of this questionnaire was to assess the users' experience and opinion of the AV product, gain insights about how the computer was used, determine the users' level of security awareness and the measure of due diligence they exert to secure their computers. Meanwhile, statistical data compiled by the scripts were collected on the computer by the experimenter. The computer was also analyzed in order to attempt finding malware missed by the AV product. The following diagnostic tools were used:

- HijackThis: gives the list of auto-loading programs and services, BHOs, IE plugins, IE toolbars, etc.;
- ProcessExplorer: shows the list of active processes;
- Autoruns: gives the complete list of programs configured to run during system bootup or login;
- Sigcheck: shows files version number, timestamp information, and digital signature details, including certificate chains;
- SpyBHORemover: gives the list of installed BHOs and classifies them in 4 categories (dangerous, suspicious, safe, unrated);
- SpyDLLRemover: gives the list of loaded DLLs and classifies them in 3 categories (dangerous, safe, unrated);
- Whatchanged: scans for modified files and registry entries;
- Winprefetchview: reads prefetch files and displays information stored in them.

We classified each element in 4 categories (safe, dangerous, suspicious, unrated) using external online resources, such as *www.systemlookup.com*, *www.processlibrary.com*, VirusTotal, Anubis and any other relevant resources. Computers with files identified as dangerous or suspicious were suspected to be infected, and any unrated files were subject to an in-depth investigation in order to confirm if they had malicious intentions. In the event that the AV product detected any malware over the course of the month, or if our diagnostic tools indicated that the laptop was infected or suspected to be, participants were asked to answer a questionnaire. This specific questionnaire aimed to collect more information regarding the potential means and sources of the infection, and on any behavioural changes observed on the computer. Moreover, additional consent was requested from the participants in order to collect specific data, such as the browser history, the tshark log files, and the suspected file(s). These data were collected to help us identify the mean and the source of the infection.

Final session The final session was similar to the other monthly sessions. However, participants had to answer a post-experiment questionnaire about their overall experience in the study. This final survey helped us identify activities or mindsets that may have unduly affected the experimental results.

We also requested that participants keep their experiment data for an additional period of 3 months in the event we might need to perform more in-depth analysis of their computer. Finally, we provided them procedures to stop the automatic collection of the data, delete the data and the tools we installed and reinstall the operating system, if they wanted to do so.

4 Lessons Learned and Recommendations

The process of designing and conducting a first pilot study of computer security clinical trial presented several challenges. We share in this section the lessons we learned and recommendations on how to better conduct computer security clinical trials in the future.

4.1 Population size estimation

Estimate the required population size before the study The required population size is based on many factors, depending of the form of the experiment, the hypotheses to be tested, the desired power of experiment results, and the budget available. While too small a population may affect the statistical significance of the results, too large a population may involve tangible costs, in time, money, and human effort. It is thus important to adequately choose the right number of participants.

Some will say that 5 users is enough to conduct usability studies [25, 15], or that 10 to 20 users are required for each condition to be tested [5]. While these guidelines can apply to qualitative studies, appropriate statistical analysis should be performed when determining the population size for quantitative studies.

The question is how many users are enough to conduct a computer security clinical trial and obtain statistically significant results? Although there is no magic number, one way to estimate the required number of subjects is to conduct a power analysis based on the desired power, the desired statistical level of significance and the effect size to be measured. The power is used as an indicator of how statistically significant the results may be. The higher the power, the higher the chance of detecting a difference between groups if it exists. Usually a power of 80% and a significance level of 5% should be considered as ideal when conducting field studies. The challenge is then to identify the desired effect size to be detected before conducting the experiment. If available, the effect size can be estimated based on prior studies or on a pilot study. When no previous study exist, the effect size can be estimated from literature review, logical assertion, and conjecture. Finally, the detailed process of the estimation of the population size should always be mentioned when reporting quantitative user studies.

For example, let us say that we want to conduct a larger-scale trial to evaluate the efficacy of an AV product. We are interested in the proportion of users that will be infected despite the fact that they are protected by a security solution. Based on the results of our previous study [9, 11], we know that 20% of the participants were infected even though they were protected, and that 38% of the total population would have been infected if they had not been protected by an AV product. A power analysis with two proportions using a Z-test can be used to estimate the population size since the outcome variable is dichotomous. Figure 1 shows that our trial should involve 99 participants for the first group and 99 participants for the control group, if we want to achieve a power of 80% with a significant level at 5%. Our trial whould then require a total population of 198 users.

4.2 Recruiting

Select users on a scientific and ethical basis Once the population size is fixed, the first step of the recruiting process is to specify the target population to be studied using appropriate inclusion and exclusion criteria. For example, one should exclude minors for ethical and legal reasons, or may decide to exclude specific type of computer users for scientific reasons. Another important exclusion is that of co-workers and friends, in order to avoid bias of the researchers and incomfortable situations of accidental disclosure of private information. During our study, we were so preoccupied to have a blind population selection protocol that it was only after the study had started that we realized that some of the subjects were indeed co-workers and acquaintances of some members of the team. Ideally the list of names should have been reviewed by team members to filter these cases, and this before the subject were assigned subject ID numbers.

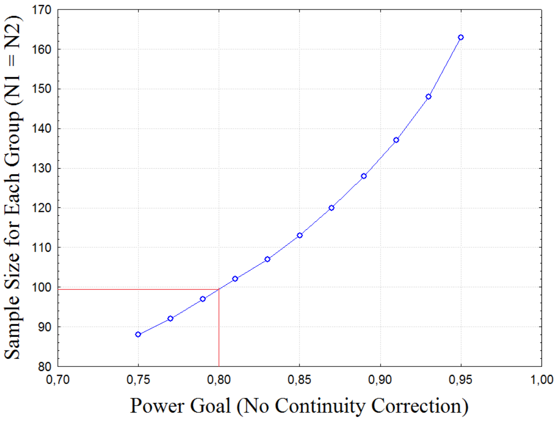

Figure 1: Population size vs. Power goal (Alpha = 5%)

Given the target population, users should be selected on their characteristics depending on the hypotheses to be tested. As mentioned, recruiting for large-scale computer security clinical trials should be structured and conducted in-house or assigned to a recruiting agency, as the "quality" of the subjects can affect the validity of results.

4.3 Data collection and validation

Identify relevant independent variables The data collected should be rigorously chosen based on the research questions. Depending on the effect to be studied, special care should be paid to the selection of the appropriate independent variables that may affect the dependent variable. One way to identify such factors is to produce an Ishikawa diagram, also known as a fishbone diagram or a cause-and-effect diagram [6]. While the common use of this causal diagram is in product design and defect prevention to identify potential factors causing an overall effect, causal diagrams can be extended to computer security.

For example, in our experiment the dependent variable to be measured is the number of malware encounters and infections per user. We identified all potential relevant factors that may influence the risk of malware infection and classified them in 5 categories (system, environment, user, usage and countermeasure) as shown in Figure 2. We then decided to fix all factors related to the system and the countermeasure, as we were interested in how the user, its behaviour and the environment affect the risk of malware encounter and infection. Unfortunately, in our case we did not follow such a deliberate approach in the initial experiment design (the Ishikawa diagram was constructed after the fact), and this led to us missing a criticial piece of information in the data collection plan. Indeed, it was only during the study that we realized how useful (and easy) it would have been to gather information about what USB media devices had been connected to the laptops during the study, as we suspected that in at least some cases they were the vehicle of infection.

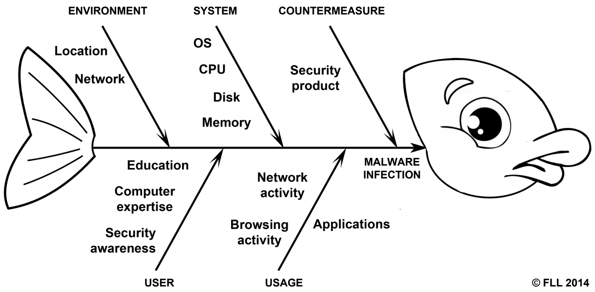

Figure 2: Fishbone diagram

Select the appropriate data collection method Once the data to be collected are identified, the next step is to chose the appropriate data collection method. The collection method should depend on the data you want to collect, but also on the research questions you want to address and the type of data analysis you want to perform.

Common collection methods can include interviews, surveys, diaries, automatic data collection, etc. The pros and the cons of each method should be considered when choosing the appropriate way to collect the data. For example, surveys are often used in user studies to collect information on users' impression, experience or behaviour. However, this method has many limitations when it comes to self-reported data because the user can either lie or report inaccurate data. One alternative to overcome this problem is to collect the same information from a reliable source in order to validate the survey results. For example, when applied to our data, we found that the reported average number of applications installed by the users was 19.02 (σ=15.17), while the real average number of applications installed was 67.78 (σ=46.66). This leads us to recommend that in the collection of self-report user behaviour data, survey results be cross-checked with another collection source before analysis.

Understand the data before the analysis In addition, prior to the analysis the collected data should be subject to an in-depth investigation for any anomalies, such as missing data or outliers. Not only it is important to identify these anomalies, but also to understand them, as they can affect the significance of the results. Appropriate statistical methods, such as residual analysis, should be used to identify and remove any potential outliers that may bias the results. For example, we were able to identify a specific user who had more than 100 detections from the AV when all other users had less than 30 detections. After investigation, we found that this high number of detections was caused by an infected file that could not be removed by the AV product, which resulted in multiple detections from the AV. We therefore decided to consider this user as an outlier and excluded it from the analysis.

4.4 Ground truthing and in-depth analysis

Our main dependent variables were whether a user had encountered or been infected by malware, how many times and with what kind of malware. One of the major difficulties that we encountered was to establish ground truth for those variables. While the number of malware encounters was provided by the AV product, the process of quantifying the number of missed detections was not so obvious. Given the fact that our tools and protocol could have failed to detect malware, the number of infections could have been underestimated. Another unexpected difficulty was that of classifying the incidents of infection. Unfortunately, malware classification is inconsistent from one AV vendor to another, and as we found, the problem is even worse with other unwanted software (adware, spyware, etc.). This is probably one of the most important weaknesses of our analysis, i.e. having to rely on inconsistent AV-provided labels to categorize the incidents.

One way to ground truth them was to conduct in-depth forensics analysis of the collected data. This was partially done in two cases, where we reverse engineered the binary executable files in order to explore the behaviour of the suspected file and make deductions about what kind it was. However, this kind of analysis is time consuming and presents many technical challenges, such as requiring significant amounts of data. For example, we did not have the complete tshark logs due to lack of disk space, which hindered our attempts at reconstructing the sequence of network events leading to infection. Because of the amount and diversity of data collected, this kind of analysis should be performed starting from clues provided by the users on what may have happened. To this purpose, we asked them to fill a questionnaire every time they had a detection or infection, in order to obtain more information on what they were doing at that time, thus helping guide the investigation and reducing its overall effort.

4.5 Bias control

Evaluate potential bias before the study Even though computer security clinical trials are less subject to bias than lab-based experiments, these biases still need to be identified and adequately controlled when possible.

In our case, we had to control system configuration variables that could affect AV performance in protecting the system. To this end, users were all using the same model of computer. In addition, they were not allowed to install another anti-malware product, neither were they allowed to uninstall or disactivate the AV product being evaluated, as these configuration options were blocked by a password. While controlling factors can reduce experimental bias, they can in return also reduce the ecological validity of the study. The challenge is then to find a suitable trade-off between bias control and result validity.

The option we choose was to evaluate the bias related to user behaviour and the environment, rather than control it. Such bias could occur when subjects share their computer with other users, use virtual machines, voluntary change their behaviour, use a private navigation mode, etc. We therefore asked the users at the end of the experiment how they may have changed their behaviour knowing that they were participating in the study.

5 Conclusion and Future Work

In previous work, we introduced a new way to evaluate security products by using the same methodology as that used in medical clinical trials. By conducting a first proof of concept study, we showed that computer security clinical trials are a viable and complementary alternative to traditional, lab-based testing of security solutions. However, designing and conducting such studies presents many scientific, technical and ethical challenges that should be carefully addressed. To this end, we have presented here the lessons learned and recommendations which would help other researchers conduct similar studies.

While our initial scientific motivation was the study and identification of user-related risk factors leading to infections, our AV industry partners were motivated by the possibility that such a methodology could be used to conduct comparative tests between different AV products, and also to assist in determining which product characteristics are more effective at protecting users. Furthermore, we believe clinical trials could also be applied to evaluate the effectiveness of training and education on users. While there have been some user studies on the impact of user education on security, most of them have been conducted in lab-based environments. From an end-user point of view and for key deciders in Information Technology (IT) management, having access to results of clinical trials will help them make better evidence-based decisions on what products and product deployment strategies constitute a more effective employment of their limited resources. Beyond choosing which product to adopt, this will include determining and addressing risky behaviour, for example by designing adequate user training and awareness sessions that is appropriately targeted to the audience, or by adopting restrictive policies. Furthermore, future clinical trials post-training will be able to measure the effectiveness and actual return on investment of such initiatives.

In summary, computer security clinical trials could improve user security by helping security product vendors improve their products, helping them understand how users adopt security solutions and react to security threats, while also helping corporations and government organizations develop better user education programs and more effective security policies.

6 Acknowledgments

This project was funded by Trend Micro and Canada's Natural Sciences and Engineering Research Council (NSERC), through the Inter-networked Systems Security Network (ISSNet) Strategic Research Network and the Discovery Grant program.

References

[1] ANTI-MALWARE TESTING STANDARDS ORGANIZATION. AMTSO suggested methods for the validation of sample.

[2] EGELMAN, S., CRANOR, L. F., AND HONG, J. You've been warned: an empirical study of the effectiveness of web browser phishing warnings. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2008), ACM, pp. 1065–1074.

[3] GARG, A., VIDYARAMAN, S., UPADHYAYA, S., AND KWIAT, K. Usim: a user behavior simulation framework for training and testing idses in gui based systems. In *Proceedings of the 39th annual Symposium on Simulation* (2006), IEEE Computer Society, pp. 196–203.

[4] HARLEY, D., AND LEE, A. Who will test the testers. In *18th Virus Bulletin International Conference* (2008), pp. 199–207.

[5] HINDERER SOVA, D., AND NIELSEN, J. How to recruit participants for usability studies, 2003.

[6] ISHIKAWA, K. *Guide to quality control*, vol. 2. Asian Productivity Organization Tokyo, 1982.

[7] KOSINAR, P., MALCHO, J., MARKO, R., , AND HARLEY, D. AV testing exposed. In *20th Virus Bulletin International Conference* (2010).

[8] Kumaraguru, P., Sheng, S., Acquisti, A., Cranor, L. F., and Hong, J. Lessons from a real world evaluation of anti-phishing training. In *eCrime Researchers Summit, 2008* (2008), IEEE, pp. 1–12.

[9] Lalonde Lévesque, F., Nsiempba, J., Fernandez, J. M., Chiasson, S., and Somayaji, A. A clinical study of risk factors related to malware infections. In *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security* (2013), ACM, pp. 97–108.

[10] Lévesque, F. L., Davis, C., Fernandez, J., Chiasson, S., and Somayaji, A. Methodology for a field study of anti-malware software. In *Workshop on Usable Security (USEC)* (2012), LNCS, pp. 80–85.

[11] Lévesque, F. L., Davis, C., Fernandez, J., and Somayaji, A. Evaluating antivirus products with field studies. In *22th Virus Bulletin International Conference* (September 2012), pp. 87–94.

[12] Lyu, M. R., and Lau, L. K. Firewall security: Policies, testing and performance evaluation. In *Computer Software and Applications Conference, 2000. COMPSAC 2000. The 24th Annual International* (2000), IEEE, pp. 116–121.

[13] Massicotte, F., Gagnon, F., Labiche, Y., Briand, L., and Couture, M. Automatic evaluation of intrusion detection systems. In *Computer Security Applications Conference, 2006. ACSAC'06. 22nd Annual* (2006), IEEE, pp. 361–370.

[14] Mell, P., Hu, V., Lippmann, R., Haines, J., and Zissman, M. An overview of issues in testing intrusion detection systems, 2003.

[15] Nielsen, J. *Usability engineering*. Elsevier, 1994.

[16] Noureldien, N. A., and Osman, I. M. On firewalls evaluation criteria. In *TENCON 2000. Proceedings* (2000), vol. 3, IEEE, pp. 104–110.

[17] PC Security Labs. Security solution review on Windows 8 platform. Tech. rep., PC Security Labs, 2013.

[18] Puketza, N., Chung, M., Olsson, R. A., and Mukherjee, B. A software platform for testing intrusion detection systems. *Software, IEEE 14*, 5 (1997), 43–51.

[19] Puketza, N. J., Zhang, K., Chung, M., Mukherjee, B., and Olsson, R. A. A methodology for testing intrusion detection systems. *Software Engineering, IEEE Transactions on 22*, 10 (1996), 719–729.

[20] Roehrig, C., and Fernandez, J. M. The supervision of research projects entailing computer risks within an academic context: The case of ecole polytechnique de montréal.

[21] Seeberg, V. E., and Petrovic, S. A new classification scheme for anonymization of real data used in ids benchmarking. In *Availability, Reliability and Security, 2007. ARES 2007. The Second International Conference on* (2007), IEEE, pp. 385–390.

[22] Shams, R., Farhan, M., Khan, S. A., and Hashmi, F. Comparing anti-spyware products - a different approach. In *Information Technology and Artificial Intelligence Conference (ITAIC), 2011 6th IEEE Joint International* (2011), vol. 1, IEEE, pp. 75–80.

[23] Somayaji, A., Li, Y., Inoue, H., Fernandez, J., and Ford, R. Evaluating security products with clinical trials. In *USENIX Workshop on Cyber Security Experimentation and Test (CSET)* (2009).

[24] Sommers, J., Yegneswaran, V., and Barford, P. Toward comprehensive traffic generation for online ids evaluation. *University of Wisconsin, Tech. Rep* (2005).

[25] Virzi, R. A. Refining the test phase of usability evaluation: how many subjects is enough? *Human Factors: The Journal of the Human Factors and Ergonomics Society 34*, 4 (1992), 457–468.

[26] Vrabec, J., and Harley, D. Real performance? In *EICAR Annual Conference* (2010).