Evaluating Security Products with Clinical Trials

Anil Somayaji
Carleton University
soma@ccsl.carleton.ca

Yiru Li
Carleton University
yiruli@ccsl.carleton.ca

Hajime Inoue
ATC-NY
hinoue@atc-nycorp.com

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

Richard Ford
Florida Insitute of Technology
rford@fit.edu

Abstract

One of the largest challenges faced by purchasers of security products is evaluating their relative merits. While customers can get reliable information on characteristics such as runtime overhead, user interface, and support quality, the actual level of protection provided by different security products is mostly unranked—or, worse yet, ranked using criteria that do not generally reflect their performance in practice. Even though researchers have been working on improving testing methodologies, given the complex interactions of users, uses, evolving threats, and different deployment environments, there are fundamental limitations on the ability of lab-based measurements to determine real world performance. To address these issues, we propose an alternative evaluation method, computer security clinical trials. In this method, security products are deployed in randomly selected subsets of targeted populations and are monitored to determine their performance in normal use. We believe that clinical trials can provide solid evidence of the efficacy of security products, much as they have in the field of medicine.

1 Introduction

The Internet is a dangerous place for users. As the reach of the network has increased, it has brought with it not only access to vast collections of data but also fraud and compromise. According to several reports [3], users are at more risk of attack than ever before. Furthermore, attackers are increasingly sophisticated, adapting quickly to new technologies and countermeasures and nimbly morphing strategies to maximize payoffs. While the security industry has mounted a valiant effort, we face a situation where our best efforts are inadequate.

Perhaps the scariest part of this situation is that we don't completely understand why we are failing. We have identifiable problems: unapplied patches, out-of-date malware signatures, poorly written software, complacent users... security experts can pontificate at length regarding the weaknesses of current systems. However, moving from this subjective, qualitative list to more concrete evaluations is difficult. Is patching *more* important than updating malware signatures? If so, how risky are delayed updates? And, more importantly, what defenses work in the field, and which ones do not? It is relatively easy to decide whether a defense could stop an attack; it is quite another to say that it will stop that attack in practice—particularly when attackers are given time to adapt and users are given the opportunity to invalidate the defense.

Today nobody knows the true relative security merits of different products, techniques, or strategies. Virus scanners perform similarly in most lab tests, with the "best" solutions differing by fractions of a percent in overall results. Firewalls are compared and sold based upon features and speed, not security. Standard security evaluation standards (such as the Common Criteria) do not apply to systems as they are used. And security experts regularly give advice such as "use strong passwords" and "turn off JavaScript" that most users will never follow. If we security experts do not know what are the best security products, and we do not know how to effectively help non-experts, is it any surprise that we have poorly secured systems?

While lab-based evaluations are essential, we believe we must do more if we are to make significant strides in improving the security of the Internet. Specifically, we must learn what works best *on deployed systems*. Note that "what works" is not the same as "what could work." For example, usability studies can identify problems that could arise in deployment, such as difficulties in firewall configuration or confusion over messages from an antivirus scanner. Ultimately, though, we don't care about usability as determined in the laboratory—we care about actual use: Do administrators misconfigure firewalls in practice? How often does user confusion over proper

virus scanner use actually lead to compromise?

To measure the use of security technologies in real-world circumstances, we have to account for how a given technology will interact with a huge variety of software, systems, users, uses, and attack profiles. The full complexity of the computational world cannot be captured in any lab setting or theoretical model—there are too many variables, and many of them change over timeframes (months or years) that cannot be practically measured in a laboratory setting using humans. As an alternative, we propose that the performance of security technologies be measured "in the field." Specifically, we propose that security technologies be tested using the same methodology as used in medical *clinical trials*. In essence, we propose that we use the same measures of outcome, side effects, and user tolerance and compliance that regulatory bodies use to demonstrate that the benefit of a drug or medical device outweighs its risks. Clinical trials come in many forms depending upon the specific questions they are designed to address; what they all have in common, though, is that the test subjects live in the "real" world, not a laboratory.

Clinical trials were originally developed because medical practitioners faced challenges analogous to those faced by today's security professionals: they knew a lot about health problems, but they didn't know what worked to prevent or fix them. Clinical trials provided a methodology for separating "snake oil" from penicillin. As we will explain, clinical trials have a number of limitations as a testing methodology; our hope, though, is that clinical trials of security technologies will allow us to separate ineffective and dangerous technologies from those that provide significant security benefits.

2 Computer Security Problems

The evaluation problem exists broadly in computer security, for both academic research and commercial products. The most egregious type of improperly evaluated security technology is often referred to as "snake oil" [8]. The ultimate question in computer security evaluation is, how do we differentiate effective security mechanisms from such quackery, particularly in the eyes of a lay audience?

Such differentiation is becoming more important because, almost always, even the best commercial systems cannot detect many of the most recent threats. This limitation arises because new threats emerge much more frequently than before, and meanwhile some of them aim for economic profits and use very complex technologies in order to bypass security mechanisms [6]. Even though many security companies have started using more flexible techniques such as heuristics to respond to new threats, in this arms race attackers always have an important advantage—the public availability of security products. Highly-skilled attackers can keep modifying their newly created malicious codes until they can bypass all current defenses [2], forcing every security vendor to constantly update their products. Given this situation, how can a regular user *know* that their vendor is providing adequate protection against the latest threats? The obvious answer is that users should check published benchmarks; unfortunately, according to those tests, virtually every major product appears to be equivalent—they all "pass" or catch virtually all tested threats.

In the antimalware field, researchers and industry members are currently working on developing better testing standards [1]; this task is extremely difficult, however, because vendors and evaluators disagree regarding basic testing practices. For example, there is no consensus on how to construct an a collection of malware for testing purposes. A major point of contention is whether such collections may contain new viruses, rather than just ones not observed "in the wild" [5].

While there are certainly ethical issues involved with creating new computer viruses, we believe there is a more fundamental issue: if you create malware from scratch for testing purposes, how do you know you've created the right kinds? In other words, how will you determine whether detection performance on synthetic test cases will correlate with performance on malware observed in practice? This issue is just one part of a much larger issue: how can you take into account all of the factors—detection mechanisms, relative frequencies of different kinds of malware, user behavior, host and network environment, changing attacker strategies and goals—that affect a product's real world performance in a set of standardized lab tests?

We believe the simple answer is that you can't—the task is impossible. There are simply too many variables. Researchers and companies will continue to argue about proper lab testing procedures because there is no single right answer: every test incorporates assumptions about the real world, and these assumptions cannot be evaluated in a laboratory setting.

Is there a way beyond this impasse? Perhaps, but only if we can test security technologies "in the field"—in the contexts in which they are used. Of course, such testing would involve attempting to protect real users from real threats while measuring relative performance. This approach is technically difficult, expensive, ethically challenging, and potentially very risky. We believe, however, that such testing is feasible based on experiences from the field of medicine, in the form of clinical trials.

3 Medical Clinical Trials

While computers and humans are very different systems, the medical field has long faced evaluation problems analogous to that of computer security. Specifically, before the 20th century there existed many potential "defenses"—treatments that promised to ensure or repair health—but people continued to be attacked and compromised (suffer and die prematurely from disease). While modern medicine has a variety of limitations, current medical practice has treatments that can reliably prevent or cure many conditions that before were debilitating or even fatal. What is remarkable about these treatments is that, in general, we don't understand how they work: our understanding of living systems is still primitive in many ways. Despite this lack of knowledge, however, we are now able to differentiate treatments that work from those that do not. The primary methodology for drawing such conclusions is the clinical trial [4].

The key insight behind clinical trials is that when studying systems (such as the human body) that are complicated, diverse, and tightly coupled with a dynamic environment, individual variables cannot be isolated and so cause and effect relationships cannot be inferred from individual observations: correlations can occur without causation, and observed effects can originate from unidentified causes. Clinical trials are an experimental methodology designed to identify causal relationships in the face of such complexity.

In medicine, clinical trials, or randomized control trials (RCTs), are planned experiments that are designed to compare treatments for a given medical condition. They use results based on a limited sample of patients to make inferences about how treatments should be conducted in the general population of patients. While the majority of clinical trials are concerned with evaluating drugs, they can also be used to evaluate other interventions such as surgical procedures, radiotherapy, physical therapy, and diets.

To account for variations in genetic makeup, lifestyle, life history, and environment, clinical trials are designed with several key features:

Selected populations At risk or afflicted individuals are studied, rather than the general population.

Extended duration Experiments are performed for months or, ideally, years in order to evaluate longer term effects.

Random samples Subjects are randomly recruited from the selected population.

Comparable Treatments Subjects are given one of a small selection of treatments, each of which is intended to treat the same condition.

Randomly Chosen Treatments Subjects or doctors do not choose their treatment; instead, the treatment is randomly assigned.

Control Groups Some subjects do not receive any treatment or are given a placebo (e.g., a sugar pill).

Blinding In a single blind study, subjects do not know which treatment they are receiving. In a double-blind study, the treating doctors do not know either.

Indicators Often the condition studied evolves over a long period of time. Rather than wait until the end (e.g., wait until the subject is cured or dead), progress is measured by observing indicators that are known to correlate with the final outcome. For example, insulin and blood sugar levels of diabetes patients are monitored in diabetes-related trials. Note that it is often hard to find a reliable indicator (e.g., a cancer recurs even when all tests indicate the treatment was successful); thus, longer term studies are always required to assess the reliability of indicators.

Due to the constraints of particular experiments, not all clinical trials will include all of these features; the more that are used, however, the greater the statistical power of the results. In other words, each of these mechanisms help with determining causal relationships. The fewer that are used, the more likely the study will only show correlation, not causation.

While clinical trials are very powerful tools for determining cause-effect relationships, they are not able to tell *why* those relationships exist. Clinical trials do not themselves provide explanations or models; what they can do, however, is test the validity and completeness of models. For example, in medicine drugs that work well in lab experiments routinely fail to work in clinical trials on people. This failure happens even when the precise molecular mechanism of the drug is known. Quite simply, we cannot capture the full complexity of the human body in any current model or lab. With clinical trials, however, we can make sure that regular patients get effective treatments—even if we don't understand how those treatments work.

4 Computer Security Clinical Trials

Because computers are engineered systems, we are much better able to determine cause and effect in computer security than in medicine. However, while it is relatively straightforward to understand a given vulnerability and devise a patch that fixes it, as we explained in Section 2, it is not nearly so easy to determine what produce the ultimate result of more secure systems. So, here we ask, is it potentially feasible to adapt the clinical trial methodology to computer security?

The key constraint to the feasibility question is to realize that clinical trials cannot be use to address the same questions as standard security evaluation techniques. We cannot use a clinical trial to analyze malware, expose a new software vulnerability, or test a new cryptographic protocol. However, we can use clinical trials to address questions such as the following:

- What is the security benefit of running an antivirus program on a personal computer in a typical home?
- Do personal firewalls provide additional protection for technically advanced users on their home machines?
- Does user training protect organizations from social engineering attacks such as phishing?

Note the key feature of these questions is that, because they involve interactions between computers and their users in specific environments, they cannot be answered in a controlled laboratory setting; nevertheless, they are precisely the kinds of questions we need to answer if we are to improve security in practice.

It takes a team of people to develop a medical clinical trial design: experts in the specific treatment must work with general clinicians, statisticians, experts in patient recruitment, ethicists, and others. Given that computer security clinical trials will also deal with human populations (along with computer populations), many of the same technical, legal, ethical, and logistical issues will need to be addressed. For these reasons, we cannot hope to present a complete trial design here; however, we can give an outline for a plausible computer security clinical trial. Here we present a sketch of a trial addressing the first question: the benefit of antivirus programs.

It is generally recommended that all personal computer users (at least, those running a version of Microsoft Windows) run an up-to-date antivirus scanner. A clinical trial designed to test their relative benefits could have the following characteristics:

Population Users running (at the start of the trial) Microsoft Windows Vista SP2 on a home machine connected to the Internet via a large home internet service provider (ISP).

Duration Three years, with preliminary results reported after each year.

Sample 1000 ISP subscribers would be randomly recruited to participate in the trial. Each subscriber would be given the following incentives to participate: free technical support and automatic offsite backups for all machines enrolled in the trial and their users. In return, they would have to agree to researchers monitoring their computer usage (subject to appropriate privacy and other controls). Users would be allowed to drop out of the trial at any time.

Treatments Three major antivirus programs would be selected for the trial and randomly assigned to different households. Note that only the given antivirus programs would be allowed to be installed; otherwise, only the standard security software that comes with Windows Vista would be allowed to be used. Compliance would be verified by scanning off-site backups.

Note that all provided software would be kept automatically up to date, including updates to the latest releases. (We assume a three year software subscription model.) Other upgrades (software and hardware) and new installations would be permitted at the user's discretion (e.g., upgrades from Windows Vista to Windows 7 and the installation of new computer games would be allowed).

Control A control group would receive no antivirus program and would be prohibited from running any host-based antivirus program. To ensure that users were still protected, unobtrusive non-host based defenses (e.g., scanning disk backups, cloud-based antivirus [7]) would need to be used. If sufficient protection could not be provided with these other mechanisms, we would then have to omit a control group. This case is analogous to a medical clinical trial where it is unethical to omit treatment for patients.

Blinding The antivirus programs would be modified to remove any obvious corporate insignia or other advertising. Color schemes would also be modified to make them as similar as possible. Otherwise, however, their interfaces would remain the same. Such uniformity would help minimize the effect a product's brand on user behavior, e.g., a new product versus a well-established brand.

In addition, if we have a control group, the control group computers would run a program that mimicked the appearance and behavior of an antivirus program. It would provide a Windows tray icon and it periodically would report that its signatures were updated. In addition, it would check and report a variety of relatively innocuous, common problems such as tracking cookies. This program would do no proper scanning and it would provide no protection from malware.

Indicators A variety of measures would be required to monitor the users and computers involved in the study. Primary measures would classify the efficacy of the tested systems based on scans of offsite backups for examples of known malware. To maximize accuracy, such scans would use a large number of commercial scanners (including those

not part of the test). Further, supplementary software would record CPU, disk, and network usage. Periodically, a small subset of machines would be inspected manually by security experts to evaluate computer health and other characteristics. Finally, technical support records would give direct measurements of time and expense.

The primary goal of such measurements would be to evaluate the "health" of the subject machines. Of course, we cannot ever be completely sure that a seemingly healthy machine is not infected. We do not need to know "ground truth" in this situation, however—we just need to measure relative performance. Thus, simplistic measures should suffice for an antivirus clinical trial.

While there are a variety of logistical, technological, and financial challenges implicit in the above description, it should be clear that it would be possible to run this trial given the right resources. While we could speculate on what results we might find from such a study, the fact is that we don't know what would be found. Indeed, that is the *key point* of clinical trials: they can reveal interactions and behaviors that are not observed in laboratories nor predicted by theoretical models.

5 Objections

There are many potential objections to the use of the clinical trial methodology in a computer security context. Here we address some of the ones that have arisen in our discussions.

5.1 Biology vs. Computers

One significant objection is that computer security is fundamentally different from medicine because the adversaries we face are not microorganisms but people—intelligent, motivated people. While many have debated the merits of the biological metaphor for computer security [9], we believe that debate is not relevant to the question of computer security clinical trials because the underlying methodology is applicable in any circumstance where one is performing experiments outside of a controlled lab setting. Randomization, selected populations, controls, blinding—these are just techniques for isolating one variable of interest from a complex background that cannot be made uniform.

Of course, it is true that clinical trials are backwards looking; thus, it is always possible that new attacks could render previously effective defenses obsolete—something that happens much less frequently in medicine. However, virtually all modern security tools also adapt to new attacks via automated update mechanisms. Thus, clinical trials of security software will, implicitly, be testing the software *and the organization behind it*. In practice, then, we would really be comparing humans (attackers) versus humans (defenders), as mediated by a computational battlefield.

But even if we are talking about human institutions, as with many financial products, past performance is not indicative of future results. Given that we cannot predict the future of security technologies using any current technique (including formal models), however, past performance is all we have to go on when choosing security solutions. Clinical trials are merely a formal methodology for rigorously assessing that past performance.

5.2 Utility

Even if adopted, a clinical trial methodology will not be a panacea with respect to security. While the approach should demonstrate the real world effectiveness of products, it will not explain *why* differences exist. For example, consider two virus scanners. Our trial would perhaps show that one product provides statistically better protection than the other—but it would not (directly) provide any explanation for their differential performance. Is it the accuracy of virus detection? The speed or ease of update? While individual users may be able to say what they liked about the product they were given, such opinions only provide clues as to the cause. As such, the results produced by the trial may be both unexpected and, *prima facie*, inexplicable.

Because of these limitations, clinical trials should be seen as a complement to, not a replacement of, lab testing of security technologies. We also believe better methodologies are needed for lab evaluations. Our purpose here, though, is to point out that lab testing cannot be expected to address all of the issues that arise in deploying security solutions. Clinical trials provide a rigorous way to determine to what extent solutions developed in the lab are applicable to practice.

5.3 Expense

To be sure, clinical trials are an expensive and complicated way to evaluate systems. Aren't there feasible alternatives? We have already discussed the limitations of lab experiments; however, there is an alternative. Rather than deal with the overhead of blinding, controls, screening populations, and the like, why not just observe real users with the defenses they already have?

Such experiments are known as observational trials. They are used frequently in medicine, particularly when researchers are searching for effects that show up over

long periods of time (e.g., decades). Unfortunately, observational trials are very limited in their ability to establish causal relationships. Thus, virtually any interesting correlation found in an observational trial is later subject to a targeted clinical trial.

While the cost of a security clinical trial can be mitigated through appropriate automation, a clinical trial will always be at least an order of magnitude more expensive than a simple lab comparison because of labor costs, particularly for technical support, subject recruitment, and ongoing observation. For example, assume that a trial required a 10:1 ratio of subjects to study personnel. Then, to run a trial with 1000 subjects we would need 100 study employees. If they are paid $100,000 on average, this study would cost $10 million/year.

We believe this estimate is a worst case scenario—effective security clinical trials should be feasible for a tenth this cost ($1,000,000/year) or less. But even this pessimistic estimate is potentially feasible: computer security is a multi-billion dollar market, and $10 million/year is well within the funding capabilities of governments or NGOs (non-profits). Further, this cost is justified by the importance of the problem. Organizations are now being required by regulation to implement security solutions. Such implementations can be very expensive. To date, we have no way of determining whether those solutions provide concrete benefits in practice.

If clinical trials are shown to work for computer security, it is likely they will become mandated by regulation, much as they have been for medicine. Such regulations would mean that changes in security practice would first need to be experimentally evaluated—for their security benefit in practice—before being adopted. We think such a change would be to the benefit of the computer security industry. Before medical practice was regulated, there was a vigorous but relatively small trade in patent medicines—unregulated preparations that claimed to cure people's ills. Despite being pioneers in marketing and advertising, patent medicines were widely maligned and mistrusted, largely because in general they didn't actually work [10]. In contrast, modern medicine is an extremely large, lucrative, and well-respected enterprise. If our community can, as a group, recommend solutions for which we have scientific evidence of their efficacy, perhaps computer security will also see a transformation in terms of its scope and prestige.

6 Conclusion

In order for the field of computer security to progress, we need better ways to measure the relative benefits of different techniques and tools as they are used in practice. To this end, we have proposed applying the proven techniques used in medical clinical trials to security. Given the importance of information assurance in the modern world and the increasing regulatory requirements for operational security, we believe the cost and complexity of clinical trials are justified. While the ultimate value of security clinical trials will only be known in retrospect, we are optimistic that clinical trials will help the development and deployment of effective security technologies.

7 Acknowledgements

The authors wish to thank Tim Furlong for first thinking of the computer security clinical trial in a lab brainstorming meeting in the summer of 2006. AS, YL, and HI acknowledge support from Canada's NSERC, though the Discovery Grants program and the Internetworked Systems Security Network (ISSNet), and MITACS.

References

[1] AMTSO. Anti-Malware Testing Standards Organization. http://www.amtso.org/.

[2] DEFCON. The Race to Zero Contest. http://www.racetozero.net/, August 8–10, 2008.

[3] FOSSI, M., Ed. *Symantec Global Internet Security Threat Report, Volume XIV*. Symantec, 2009.

[4] FRIEDMAN, L. M., FURBERG, C. D., AND DEMETS, D. L. *Fundamentals of Clinical Trials*, 3rd ed. Springer-Verlag, 1998.

[5] KREBS, B. Anti-virus testing and consumer reports. *Security Fix* (August 29, 2006). http://voices.washingtonpost.com/securityfix/2006/08/antivirus_testing_and_consumer_1.html.

[6] LARKIN, E. Storm Worm's virulence may change tactics. *Network World* (August 2, 2007).

[7] OBERHEIDE, J., COOKE, E., AND JAHANIAN, F. CloudAV: N-Version Antivirus in the Network Cloud. In *17th USENIX Security Symposium* (2008).

[8] SCHNEIER, B. Snake oil. *Crypto-Gram Newsletter* (February 15, 1999). http://schneier.com.

[9] SOMAYAJI, A., LOCASTO, M., AND FEYEREISL, J. Panel: The Future of Biologically-Inspired Security: Is There Anything Left to Learn? In *2007 Workshop on New Security* (2008), ACM.

[10] STYLES, J. Product Innovation in Early Modern London. *Past & Present 168*, 1 (2000), 124–169.