Citation: A. G. Dunn, R. O. Day, K. D. Mandl, E. Coiera, Learning from hackers: Open source clinical trials. *Sci. Transl. Med.* **4**, 132fs10 (2012).

Learning from Hackers: Open-Source Clinical Trials

Adam G. Dunn,[1]* Richard O. Day,[2] Kenneth D. Mandl,[3,4] Enrico Coiera[1]

[1]Centre for Health Informatics, Australian Institute of Health Innovation, University of New South Wales, Sydney, New South Wales 2052, Australia.

[2]St Vincent's Clinical School and School of Medical Sciences, University of New South Wales, Sydney 2052, NSW, Australia.

[3]Children's Hospital Informatics Program at the Harvard-MIT Division of Health Sciences and Technology, Children's Hospital Boston, Boston, MA 02115, USA.

[4]Department of Pediatrics, Harvard Medical School, Boston, MA 02115, USA.
*Corresponding author. E-mail: a.dunn@unsw.edu.au

The open source software movement can serve as a model for a similar initiative in the clinical trial community.

ABSTRACT

Open sharing of clinical trial data has been proposed as a way to address the gap between the production of clinical evidence and the decision-making of physicians. A similar gap was addressed in the software industry by their open-source software movement. Here, we examine how the social and technical principles of the movement can guide the growth of an open-source clinical trial community.

HACKERS AS ROLE MODELS

Despite the rapid increase in the volume of published biomedical research (*1*), physicians often make decisions without access to the synthesized evidence they need to practice evidence-based medicine (*2*, *3*). Although improvements in research translation have been made through mandatory clinical trial registration, patient advocacy, and public-private partnerships (*4*–*9*), the gap between biomedical advances and their incorporation into clinical practice remains a grand challenge for health care (*10*). Hackers—authority-questioning software developers who helped orchestrate the free software movement—successfully addressed a similar gap in the software industry. This iconoclastic group fundamentally altered the software industry by devising the principles that drive the open-source software movement, countering the economic and cultural motivations that drove the production of closed-source software, disengagement with user needs, and poor interoperability. Similar roadblocks plague the clinical evidence domain. Here, we describe how the open-source software movement can guide growth of an emerging open source effort in the clinical trial community (Fig. 1).

FREEDOM OF INFORMATION

The open-source software movement grew out of the intellectual curiosity of hackers and a fundamental belief that all information should be free (*11*). In this case, "free" meant "libre" (no restrictions) rather than just "gratis" (zero cost). The so-called four freedoms of open-source software include the freedom to run a program for any purpose, to study how the program works and change it, to redistribute copies, and to distribute modified copies (*12*).

Success of the open-source software movement is driven by the developers' active engagement with users in the software creation and testing phases, and the rapid filling of gaps in functionality by a decentralized cadre of developers, a process supported by online communities formed around repositories of code. GitHub, the largest code host in the world, was started in 2008 and has 1.2 million users and hosts 3.5 million repositories. Source Forge is the second largest and most established, with over 326,000 diverse software projects supported by a globally distributed community of 2.7 million developers. Examples of widely used open-source software include the Firefox Web browser, the Apache Web server, and the Android and Symbian operating systems, which are used globally on smart phones.

Open source communities often out-perform their closed source counterparts when addressing the needs of users (*13*), arguably as a direct consequence of the dialogue with users and the interoperability that comes from transparency and developer interaction (*14*). Users of open-source software are encouraged to become involved directly in the process by reporting issues and gaps in functionality that they wish to see addressed, even if they are unable to contribute directly to software development.

Individuals and organizations of all sizes participate in open-source software communities with a variety of motivations that range from career development to reputation building. Large companies find value in open-source software development as a way to improve the "social contagion" of their products by engaging directly with heavy users (*15*). Industry giants profit by offering complementary services (*16*) and engaging with open source communities to identify talent and recruit open source developers to their companies (for an online curriculum vitae service, see http://geekli.st/beta).

FREEING CLINICAL DATA

The process of translating clinical evidence into practice is slowed by limited sharing of patient-level clinical trial data and bottlenecks in the dissemination of evidence. Industry often seeks to answer questions in clinical trials that relate to their business agendas rather than answer the questions of most value to clinicians and patients (*10, 17*). These systemic problems have been attributed to biases that affect which clinical trials are conducted and which results are published (*18*). The slow dissemination of trustworthy evidence can be obscured by industry marketing delivered through the same channels and the conflicts of interest and low-level evidence that permeates clinical practice guidelines (*19*).

Notable successes have come from the expanding of access to de-identified patient-level data. The Framingham Heart Study resulted in 1872 publications between 1948 and 2007, and when access to genotypic and phenotypic data was opened to a much wider group of researchers, this number increased by 19% to 2223 in just 4 years (*20*). In a separate example, the public release of genotypic information associated with an outbreak of *Escherichia coli* in May 2011 (40 deaths and 3000 cases of infection) led to genomic analyses on four continents (*21*). Within a week of public release, the bacterial genome had been assembled and the strain identified, and two dozen reports had been filed that provided information on bacterial strain virulence, drug resistance, and phylogenetic lineage. The speedy release of the reports on the *E. coli* outbreak improved clinical practice by indicating

which drugs to use (or not use) to treat patients and may have contributed to identification of the source of the outbreak by examining the lineage. In another example, a patient-level analysis of a cholesterol treatment was produced collaboratively by researchers who collectively had conducted 26 clinical trials that included a total of ~170,000 patients (*22*). The protocol for the patient-level meta-analysis was agreed on and published in 1995 before any of the individual trials were completed, making this an early example of broad collaboration and prospective sharing of patient-level clinical trial data.

There are strong similarities between the processes for collaborative engagement with software source code and patient-level data from clinical trials. There are differences, too, such as the expertise and infrastructure required when producing source code and clinical trial data, privacy concerns associated with patient identification, and the sources of funding that underpin the two systems.

EXTRAPOLATING TO PRINCIPLES

The analogy between open-source software development and the synthesis of clinical evidence suggests the following principles for an open-source clinical trial community:

(i) Clinical trialists, other scientific researchers, and clinicians can access and contribute to a repository of interoperable patient-level data that sits alongside the mandatory registration and provision of summary-level clinical trial results.

(ii) Infrastructure for repositories, data standards, and interaction among clinical trialists must be provided to foster the growth of a decentralized community whose main aim is to rapidly identify and address gaps in clinical evidence.

(iii) Improved dialogue between trialists and physicians, which would allow physicians to routinely ask new questions of the clinical trial community and follow and discuss the aggregated clinical trial data to answer existing questions.

An open-source clinical trial community requires data standards that allow trials to be recorded in a uniform way while retaining the flexibility to represent the variety of protocols and interventions that exist in current clinical trial registries (*23*). The development of data standards can be slow and laborious, but this is an area in which much work has been done (*24*). Sim and colleagues (*25*) have led the design of data standards to capture machine-interpretable knowledge bases and the design of a framework for storing summary information with enough detail for searching and aggregating across clinical trials. Other examples of standards and developing communities that support the exchange of clinical trial data include the Study Data Tabulation Model (*26*) and Open mHealth (*27*). By leveraging the most flexible of the data standards, it should be possible to create, store, and aggregate data sets without losing the rigor of the protocols established within individual clinical trials.

We envision a set of tools similar to those provided by GitHub or SourceForge, which will allow clinical trial researchers to submit deidentified patient-level outcome data alongside the usual metadata expected when registering clinical trials (*23*). Each submitted module would represent patient-level outcomes separated by study arms, coupled with information about the patient inclusion criteria and the interventions applied. Modules with equivalent interventions may be combined and compared against alternatives to produce the equivalent of patient-level meta-analyses. Outcome differences between groups of patients who received the same intervention can still be analyzed against differences in phenotype, genotype, and setting (combining homogeneous groups and performing subgroup analysis between heterogeneous groups), just as is today done in multicenter trials and meta-analyses. The ultimate benefit of such an approach would be an increase in the utility of all clinical

trials through improved connectivity between, and access to, data needed for patient-level meta-analyses; such collaborations and analyses may thus permit construction of a richer evidence base for clinical practice guidelines.

Patient-level meta-analyses are rarely performed. Only ~50 patient-level meta-analyses are published each year, compared with the ~17,500 clinical trials registered with clinicaltrials.gov, and the 75 clinical trials and 11 systematic reviews published each day (*1*, *23*). This is despite a long list of advantages that patient-level meta-analyses have over meta-analyses based on trial summary information (*28*). The disadvantage of patient-level meta-analysis is the time-consuming nature of coordination and data management issues that would be largely resolved by standardizing data storage methods, information reporting standards, and avenues for worldwide collaboration and communication. As a consequence of sharing, it will also be possible to genuinely decouple the collection of data and the analysis for study types ranging from controlled trials of single interventions to more complex studies such as network meta-analyses. Under most open source licenses, attribution of published syntheses would automatically flow back to the individuals and groups who contribute data, and it would be possible to extend open source licenses to include coauthorship privileges.

As is the case for software users in the open-source software movement, the open-source clinical trial community would facilitate the direct participation of physicians who are seeking decision support. As an open-access repository, physicians would have access to the conclusions drawn from large patient-level meta-analyses and find trust in the transparency of the analyses and underlying data. Physicians could request answers to new and clinically relevant questions that have not yet been sufficiently addressed by clinical trials (*29, 30*). This could be done in the same way that software features are requested in open source communities—via an online submission system. As a consequence, a direct dialogue is created between those producing the evidence and the physicians using the evidence in their decision-making process in the care of patients.

By creating a platform to improve the way patient-level data are shared and directly engaging the users of clinical evidence in the process, the data developers (clinical and translational scientists) will be better able to provide the evidence that physicians need in their practices. Researchers also will improve the utility of the evidence they produce by reusing outcome data in subsequent analyses and reducing the bottlenecks associated with data access (*31*). These enhancements can be expected to improve the utility of clinical trial data through reuse and more appropriate aggregation and synthesis and reduce the burden of waste associated with research data more generally (*32*).

CONFRONTING THE CHALLENGES

Technical challenges around building an open source community for clinical trials include the application of methods to avoid or eliminate the potential for identifying individual patients (*33*) and implementation of new ways to address data quality standards, which can be low for decentralized contributions (*4*, *18*). Less onerous challenges include providing the software infrastructure and tools that foster discussion, growth, and dialogue with physicians and tools that aid the self-governance of quality.

Privacy is clearly a concern when providing open access to patient-level data for cases in which deidentification cannot be ensured (*33*). One solution is to require signed agreements to not release information that could be used to identify individuals, as is currently done for access to existing longitudinal data sets. An alternative would be to generate statistically identical samples on request, allowing all necessary inferences to be made without compromising privacy.

Perhaps the most important challenge faced by the clinical trial community comes from the currently unbalanced value system in which publish-or-perish mentalities and marketing concerns often outweigh the value of making effective contributions to the support of clinical decision-making. Pharmaceutical companies may have an aversion to providing open access to patient-level data because it may reduce their ability to control the conclusions that are drawn or to avoid dissemination of unfavorable results. In open-source software the incentives for participation are understood, even by large companies (*16*, *34*), and these values may also be discovered by pharmaceutical companies. Addressing the participation challenge will depend on the choosing or creating of licenses that usefully capture contributions in order to confer recognition for those contributions in publications and by career progression. Software developers routinely use open-source contributions to compete for career advancement.

There has been a recent international shift toward requiring open access to all publications that result from publicly funded clinical trials. A further push—at least for trials with nonindustry funding sources—may come from extending the mandate of open-access publication to include open access to patient-level data for publicly funded clinical trials.

Despite the technical and social challenges, substantial movement toward crowdsourcing and open access to data has already been seen in early-phase drug development (*7*, *35*, *36*), apparently in response to the slowing of approvals that signaled a grand challenge for the field (*37*). The domain of clinical data for the practice of evidence-based medicine faces its own grand challenge and requires a similar push to close the gap between what physicians need and the ways in which trials are funded, conducted, and reported. By recognizing the open-source software community as a role model for improvement, the clinical and translational research community can establish principles, standards, and tools that catalyze the growth of a more efficient and socially responsible clinical trial community.

Fig. 1. Hacking away at the obstacles. Through open source communities, such as the one facilitated by SourceForge, the software industry has resolved bottlenecks similar to those impeding the translation of clinical trial evidence into medical practice .

CREDIT: B. STRAUCH/*SCIENCE TRANSLATIONAL MEDICINE*

References and Notes

<jrn>1. H. Bastian, P. Glasziou, I. Chalmers, Seventy-five trials and eleven systematic reviews a day: How will we ever keep up? *PLoS Med.* **7**, e1000326 (2010). doi:10.1371/journal.pmed.1000326 Medline</jrn>

<jrn>2. T. A. Kotchen, Why the slow diffusion of treatment guidelines into clinical practice? *Arch. Intern. Med.* **167**, 2394–2395 (2007). doi:10.1001/archinte.167.22.2394 Medline</jrn>

<jrn>3. F. Davidoff, J. Miglus, Delivering clinical evidence where it's needed: Building an information system worthy of the profession. *JAMA* **305**, 1906–1907 (2011). doi:10.1001/jama.2011.619 Medline</jrn>

<jrn>4. K. Dickersin, D. Rennie, Registering clinical trials. *JAMA* **290**, 516–523 (2003). doi:10.1001/jama.290.4.516 Medline</jrn>

<jrn>5. N. Rasmussen, K. Lee, L. Bero, Association of trial registration with the results and conclusions of published trials of new oncology drugs. *Trials* **10**, 116 (2009). doi:10.1186/1745-6215-10-116 Medline</jrn>

<jrn>6. K. Lutchen, J. Ayers, S. Gallagher, L. Abu-Taleb, Engineering efficient technology transfer. *Sci. Transl. Med.* **3**, 110cm32 (2011).</jrn>

<jrn>7. B. H. Munos, W. W. Chin, How to revive breakthrough innovation in the pharmaceutical industry. *Sci. Transl. Med.* **3**, 89cm16 (2011). doi:10.1126/scitranslmed.3002273 Medline</jrn>

<jrn>8. T. C. Norman, C. Bountra, A. M. Edwards, K. R. Yamamoto, S. H. Friend, Leveraging crowdsourcing to facilitate the discovery of new medicines. *Sci. Trans. Med.* **3**, 88mr1 (2011).</jrn>

<jrn>9. S. F. Terry, P. F. Terry, Power to the people: Participant ownership of clinical trial data. *Sci. Transl. Med.* **3**, 69cm3 (2011).</jrn>

<jrn>10. I. Chalmers, P. Glasziou, Avoidable waste in the production and reporting of research evidence. *Lancet* **374**, 86–89 (2009). doi:10.1016/S0140-6736(09)60329-9 Medline</jrn>

<bok>11. S. Levy, *Hackers: Heroes of the Computer Revolution*. (Anchor Press/Doubleday, Garden City, New York, 1984), pp. 458.</bok>

<eref>12. The Free Software Foundation, The Free Software Definition, *Accessed 14 July 2011*, http://www.gnu.org/philosophy/free (1996).</eref>

<jrn>13. J. W. Paulson, G. Succi, A. Eberlein, An empirical study of open-source and closed-source software products. *IEEE Trans. Softw. Eng.* **30**, 246–256 (2004). doi:10.1109/TSE.2004.1274044</jrn>

<jrn>14. E. Raymond, The cathedral and the bazaar. *Knowl. Tech. Policy* **12**, 23–49 (1999). doi:10.1007/s12130-999-1026-0</jrn>

<jrn>15. R. Iyengar, C. Van den Bulte, T. W. Valente, Opinion leadership and social contagion in new product diffusion. *Mark. Sci.* **30**, 195–212 (2011). doi:10.1287/mksc.1100.0566</jrn>

<jrn>16. A. Bonaccorsi, C. Rossi, Comparing motivations of individual programmers and firms to take part in the open source movement: From community to business. *Knowl. Tech. Policy* **18**, 40–64 (2006). doi:10.1007/s12130-006-1003-9</jrn>

<jrn>17. A. G. Dunn, F. T. Bourgeois, S. Murthy, K. D. Mandl, R. O. Day, E. Coiera, The role and impact of research agendas on the comparative-effectiveness research among antihyperlipidemics. *Clin. Pharmacol. Ther.* **91**, 685–691 (2012). doi:10.1038/clpt.2011.279 Medline</jrn>

<jrn>18. F. T. Bourgeois, S. Murthy, K. D. Mandl, Outcome reporting among drug trials registered in ClinicalTrials.gov. *Ann. Intern. Med.* **153**, 158–166 (2010). Medline</jrn>

<jrn>19. M. D. Cabana, C. S. Rand, N. R. Powe, A. W. Wu, M. H. Wilson, P. A. Abboud, H. R. Rubin, Why don't physicians follow clinical practice guidelines? A framework for improvement. *JAMA* **282**, 1458–1465 (1999). doi:10.1001/jama.282.15.1458 Medline</jrn>

<eref>20. Framingham Heart Study, Framingham Heart Study Bibliography. *Accessed 14 July 2011*, http://www.framinghamheartstudy.org/biblio/index.html (2011).</eref>

<jrn>21. H. Rohde, J. Qin, Y. Cui, D. Li, N. J. Loman, M. Hentschke, W. Chen, F. Pu, Y. Peng, J. Li, F. Xi, S. Li, Y. Li, Z. Zhang, X. Yang, M. Zhao, P. Wang, Y. Guan, Z. Cen, X. Zhao, M. Christner, R. Kobbe, S. Loos, J. Oh, L. Yang, A. Danchin, G. F.

Gao, Y. Song, Y. Li, H. Yang, J. Wang, J. Xu, M. J. Pallen, J. Wang, M. Aepfelbacher, R. Yang, *E. coli* O104:H4 Genome Analysis Crowd-Sourcing Consortium, Open-source genomic analysis of Shiga-toxin-producing *E. coli* O104:H4. *N. Engl. J. Med.* **365**, 718–724 (2011). doi:10.1056/NEJMoa1107643 Medline</jrn>

<jrn>22. Cholesterol Treatment Trialists' (CTT) Collaboration, Efficacy and safety of more intensive lowering of LDL cholesterol: A meta-analysis of data from 170,000 participants in 26 randomised trials. *Lancet* **376**, 1670–1681 (2010). doi:10.1016/S0140-6736(10)61350-5 Medline</jrn>

<jrn>23. D. A. Zarin, T. Tse, R. J. Williams, R. M. Califf, N. C. Ide, The ClinicalTrials.gov results database—Update and key issues. *N. Engl. J. Med.* **364**, 852–860 (2011). doi:10.1056/NEJMsa1012065 Medline</jrn>

<jrn>24. E. Coiera, Building a national health IT system from the middle out. *J. Am. Med. Inform. Assoc.* **16**, 271–273 (2009). doi:10.1197/jamia.M3183 Medline</jrn>

<jrn>25. I. Sim, D. E. Detmer, Beyond trial registration: A global trial bank for clinical trial reporting. *PLoS Med.* **2**, e365 (2005). doi:10.1371/journal.pmed.0020365 Medline</jrn>

<jrn>26. F. Wood, T. Guinter, Evolution and implementation of the CDros. Inf. Serv.C Study Data Tabulation Model (SDTM). *Pharmaceutical Programming* **1**, 20–27 (2008). doi:10.1179/175709208X334623</jrn>

<jrn>27. D. Estrin, I. Sim, Health care delivery. Open mHealth architecture: An engine for health care innovation. *Science* **330**, 759–760 (2010). doi:10.1126/science.1196187 Medline</jrn>

<jrn>28. R. D. Riley, P. C. Lambert, G. Abo-Zaid, Meta-analysis of individual participant data: rationale, conduct, and reporting. *BMJ* **340**, (feb05 1), c221 (2010). doi:10.1136/bmj.c221 Medline</jrn>

<jrn>29. A. Mullard, The long Avandia endgame. *Lancet* **378**, 113 (2011). doi:10.1016/S0140-6736(11)61071-4 Medline</jrn>

<jrn>30. N. H. Goldberg, S. Schneeweiss, M. K. Kowal, J. J. Gagne, Availability of comparative efficacy data at the time of drug approval in the United States. *JAMA* **305**, 1786–1789 (2011). doi:10.1001/jama.2011.539 Medline</jrn>

<jrn>31. P. C. Gøtzsche, A. W. Jørgensen, Opening up data at the European Medicines Agency. *BMJ* **342**, d2686 (2011). doi:10.1136/bmj.d2686 Medline</jrn>

<jrn>32. B. Hanson, A. Sugden, B. Alberts, Making data maximally available. *Science* **331**, 649 (2011). doi:10.1126/science.1203354 Medline</jrn>

<jrn>33. G. Loukides, J. C. Denny, B. Malin, The disclosure of diagnosis codes can breach research participants' privacy. *J. Am. Med. Inform. Assoc.* **17**, 322–327 (2010). Medline</jrn>

<jrn>34. J. Lerner, J. Tirole, Some simple economics of open source. *J. Ind. Econ.* **50**, 197–234 (2002). doi:10.1111/1467-6451.00174</jrn>

<jrn>35. M. S. Boguski, K. D. Mandl, V. P. Sukhatme, Repurposing with a difference. *Science* **324**, 1394–1395 (2009). doi:10.1126/science.1169920 Medline</jrn>

<jrn>36. B. H. Munos, W. W. Chin, A call for sharing: Adapting pharmaceutical research to new realities. *Sci. Transl. Med.* **1**, 9cm8 (2009).</jrn>

<jrn>37. S. M. Paul, D. S. Mytelka, C. T. Dunwiddie, C. C. Persinger, B. H. Munos, S. R. Lindborg, A. L. Schacht, How to improve R&D productivity: The pharmaceutical industry's grand challenge. *Nat. Rev. Drug Discov.* **9**, 203–214 (2010). Medline</jrn>

Acknowledgments: The authors acknowledge the valuable comments of four anonymous reviewers. **Funding**: Supported by the National Health and Medical Research Council (program grant 568612) and 1R01LM011185-01 from the National Library of Medicine. **Competing interests:** The authors declare no competing interests.

Citation: A. G. Dunn, R. O. Day, K. D. Mandl, E. Coiera, Learning from hackers: Open-source clinical trials. *Sci. Transl. Med.* **4**, 132fs10 (2012).