Accountability, Secrecy, and Innovation in AI-Enabled Clinical Decision Software

(preprint, final version to appear in *Journal of Law and the Biosciences*, volume 7, issue 2)

Authors: Arti K. Rai (corresponding author)
Duke Law School
210 Science Drive
Durham, NC 27705
(919) 613-7276
rai@law.duke.edu

Isha Sharma
Duke-Margolis Center for Health Policy

Christina Silcox
Duke-Margolis Center for Health Policy

Acknowledgments: This work was supported by the Greenwall Foundation. We thank Michael Abramoff, Pat Baird, Kathy Blake, Susan Dentzer, Barbara Evans, Richard Frank, Kate Gaudry, Erich Huang, Mo Kaushal, Ziad Obermeyer, Nicholson Price, Guillermo Sapiro, and David Vidal for their invaluable insights. Michael McCarty, Meredith Stewart, and Bennett Wright provided excellent research assistance.

Abstract:

Employing analytical and empirical tools, this article investigates the intricate interrelationship of trade secrecy, accountability, and innovation incentives in clinical decision software enabled by machine learning (ML-CD). While trade secrecy can provide incentives for innovation, it can also diminish the ability of third parties to adjudicate risk and benefit responsibly. However, the type of information FDA and adopters are asking for, and developers are willing to provide, itself represents something of a black box.

Our article shines light into the black box. We find that developers regard secrecy over training data and details of the trained model as central to competitive advantage. Some believe uncertainty regarding the availability and enforceability of patents in ML-CD contributes to secrecy. Meanwhile, neither FDA nor adopters are currently asking for these types of details. Additionally, in some cases, it is not clear whether developers are being asked to provide rigorous evidence of performance. FDA, developers, and adopters could all do more to promote information flow, particularly as ML-CD models move into areas of higher risk. Specifically, FDA and developers could, without sacrificing innovation incentives, release summary information regarding training data and process. Moreover, particularly for higher risk cases, FDA and adopters could ask for evidence of performance on completely independent data sets. Consistent with protecting trade secrecy, FDA could also set up procedures by which to ask for full details regarding data and models.

Key terms: machine learning; clinical decision software; accountability; secrecy; innovation; intellectual property

Accountability, Secrecy, and Innovation Incentives in AI-Enabled Clinical Decision Software

INTRODUCTION

Artificial intelligence (AI) has established a substantial foothold in health care, including clinical decision making,[1] and has the potential to provide better clinical results less expensively. At the same time, AI models based on unsuitable data have been shown to reach decisions that are clinically inaccurate, with these inaccuracies sometimes systematically directed to historically vulnerable groups.[2] To avoid unintended harms, actors in the development and adoption ecosystem must promote accountability. Accountability starts with procedures that

[1] Eric J. Topol, *High Performance Medicine: The Convergence of Human and Artificial Intelligence*, 29 NATURE MEDICINE 44 (2019).
[2] Ziad Obermeyer et al., *Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations*, 366 SCIENCE 447 (2019).

assure careful evaluation of risk and benefit relative to plausible alternatives, such as reliance on human decision making.

In this Article, we examine the current state of accountability in AI-enabled clinical decision software. We use the term clinical decision software broadly, to encompass tools that span the spectrum from informing the clinician to treating or diagnosing.[3] We focus specifically on accountability issues raised by the subset of AI-based clinical decision making that is supervised machine learning (hereinafter ML-CD). Currently, most ML-CD methods that achieve human-level performance in clinical decision making rely on supervised learning.[4]

Supervised machine learning is a procedure in which the data scientist exposes a learning algorithm to "training data" that experts in the field have identified as having potentially relevant input features and have then labeled with respect to desired output classification. For example, a data scientist could give a learning algorithm a set of mammograms, of which some percentage had been labeled by one or more reviewing radiologists as having breast cancer.[5] Assuming the radiologists had done their job properly, this dataset should represent something close to "ground truth."[6] Data scientists would then subdivide the ground truth data set into a training set, a tuning set, and a test set. They would use the training set to directly learn parameters for the model and then use the tuning set to choose hyperparameters (e.g. learning rate) that affect the final

[3] We use the term clinical decision software instead of the more common term "clinical decision support" (CDS) software because we include in our inquiry software as a medical device (SaMD) that goes beyond supporting clinicians in their decision making by driving or automating the next medical intervention. .
[4] Yun Liu et al., *How To Read Articles that Use Machine Learning: Users' Guides to the Medical Literature*, 322 JAMA 1806, 1808 (2019).
[5] This hypothetical example is taken from W. Nicholson Price II and Arti K. Rai, *Clearing Opacity Through Machine Learning*, 106 IOWA L. REV. __ (forthcoming 2021)
[6] *See generally* RICHARD BERK, STATISTICAL LEARNING FROM A REGRESSION PERSPECTIVE (2nd ed. 2016).

derived parameters.[7] Once predictive performance met expectations, they would test the final trained model on the test set. Ideally, for purposes of ensuring generalizability, they would also test the model on a completely independent data set. The model that emerged could either be locked, or it could be allowed to learn in real-time from new data to which it was exposed.

ML-CD models raise challenges for responsible evaluation of risk and benefit. Some of these challenges are not unique to ML-CD, but can arise with any predictive algorithm. For example, as commentators have noted, algorithm designers can err with respect to choice of relevant endpoint or dependent variable, can fail to benchmark against standard of care, and can fail to test for generalizability.[8]

In contrast, one feature of ML-CD models that distinguishes them from other predictive algorithms is the potential for substantial opacity. Although ML-CD models exist on a spectrum, and generalizations are therefore tricky, at least some models can be opaque, even to experts who have a background in both relevant biology and data science. In many cases, this is because the models are quite complex in terms of the number of interdependent input factors they evaluate.[9] Machine learning models can also be non-intuitive – experts may be able to observe the decision rules, but the rules may not make sense to the experts.[10]

---

[7] See Zech, J. et al., Confounding variables can degrade generalization performance of radiological deep learning models, available at https://arxiv.org/abs/1807.00431 In addition to the parameters upon which all statistical models rely, machine learning models also rely on hyperparameters.

[8] *See, e.g.* Obermeyer et al., *supra* note __ (discussing model that aimed to predict future health needs but used as its dependent variable future health care expenditures, with the result that racial minorities that underutilize health care relative to need were disadvantaged); Ravi Parikh et al., *Regulation of Predictive Analytics in Medicine: Algorithms Must Meet Regulatory Standards of Clinical Benefit*, 363 SCIENCE 810 (2019) (discussing problems with respect to meaningful endpoints, appropriate benchmarks, and generalizability, along with other concerns).

[9] *See* Jenna Burrell, *How the Machine "Thinks": Understanding Opacity in Machine Learning Algorithms*, 3 BIG DATA & SOC'Y 1 (2016); *see also* Andrew D. Selbst & Solon Barocas, *The Intuitive Appeal of Explainable Machines*, 87 FORDHAM L. REV. 1085, 1095 (2018) (using the term "inscrutability" to refer to the "situation in which the rules that govern decision-making are so complex, numerous, and interdependent that they defy practical inspection and resist comprehension.")

[10] *See* Burrell *supra* ___.

Opacity created by complexity and non-intuitiveness is often exacerbated by deliberate secrecy that results from the desire of commercial developers to protect their innovation from copying by competitors.

For several reasons, our Article's inquiry into opacity and accountability focuses on secrecy. First, unlike complexity and non-intuitiveness, secrecy is an institutional behavior directly amenable to legal levers for improving information flow.

Second, addressing secrecy presents a legally and intellectually complex puzzle. Although secrecy can pose challenges for accountability, the empirical evidence indicates that it can also serve as an important incentive for innovation.[11] In the case of ML-CD, the complexity and non-intuitiveness of certain types of machine learning may enhance the competitive advantage conferred by secrecy. The logic runs as follows: under boilerplate intellectual property law, legal regimes governing secrecy offer protection to decision making models only to the extent the model cannot be reverse engineered through information lawfully available to a competitor.[12] Like software generally, ML-CD models are often cloud-based. Although computer scientists are doing work on reverse engineering cloud-based machine-learning models, this work gets more difficult, and potentially easier for developers to detect and guard against, as the model gets more complex.[13]

Third, just as technological advance has put a spotlight on secrecy, so have some recent legal changes. These include the Supreme Court's decisions in the 2012 case *Mayo v.*

---

[11] *See infra* notes 30-31 and accompanying text.

[12] *See generally* Pamela Samuelson and Suzanne Scotchmer, *The Law and Economics of Reverse Engineering*. 111 YALE LAW JOURNAL 1575 (2001).

[13] *See, e.g.*, Seong Joon Oh et al., *Towards Reverse-Engineering Black-Box Neural Networks*, available at https://arxiv.org/abs/1711.01768; Florian Tramer et al., *Stealing Machine Learning Models via Prediction APIs*, available at https://arxiv.org/abs/1609.02943.

Prometheus[14] and the 2014 case *CLS Bank v. Alice*,[15] respectively indicating that medical decision making methods that focus on diagnosis may represent a patent-ineligible law of nature and that software may represent a patent-ineligible abstract idea. ML-CD software lies at the nexus of the types of technology targeted by *Mayo* and *Alice*.

In part, these decisions were a response to concerns that the U.S. Patent and Trademark Office (USPTO) was granting low-quality patents, particularly in software, that were then being used in lawsuits against genuine innovators.[16] However, because patents provide a form of protection that survives public disclosure, and can therefore provide an alternative to secrecy, the decisions may have enhanced the attractiveness of secrecy.[17] Additionally, Congressional passage of the 2016 Defend Trade Secrets Act,[18] which provides a uniform national framework for pursuing violations of contractual obligations to keep information secret, not only reflects increasing national attention to secrecy but may further bolster its value.

Finally, concerns over health data privacy may have enhanced the value of secrecy. Secrecy may be cheaper than other mechanisms for complying with privacy requirements. Additionally, relative to competitive advantage, privacy may provide a more publicly acceptable reason for justifying secrecy. Thus, for example, to the extent developers did not want to release training data publicly for competitive reasons, they could justify this stance on the ground that public release of even de-identified training data might allow re-identification[19]

Mediating tensions between secrecy and accountability while at the same time

---

[14] 566 U.S. 66 (2012).
[15] 573 U.S. 208 (2014).
[16] *See generally* Paul Gugliuzza, *Quick Decisions in Patent Cases*, 106 GEORGETOWN L.J. 619 (2017).
[17] *See, e.g.*, Christi Guerrini et al., *Constraints on Gene Patent Protection Fuel Secrecy Concerns: A Qualitative Study*, 4 JOURNAL OF LAW AND THE BIOSCIENCES 542 (2017).
[18] Pub. L. 114–153 (2016).
[19] The limited success of legislation like the HITECH Act in achieving interoperability of electronic health records many also discourage movement away from the default of secrecy-protected siloes.

maintaining innovation incentives will require careful calibration of information flow. For software that falls within the regulatory jurisdiction of the FDA, such regulation represents an important calibration mechanism. The actions of private sector developers and adopters may also play a role, particularly for software that is not currently regulated. And for both regulated and unregulated software, tort liability regimes should be structured in a manner that shapes appropriate information flow. However, the type of information FDA and adopters are asking for, and developers are willing to provide, itself represents something of a black box. This Article provides an early analytical and empirical examination of information flow between developers, FDA, and adopters.

A note on scope: although ML-CD models are capable of updating automatically based on exposure to new data, we focus on locked models. Models that update automatically not only raise a plethora of somewhat speculative challenges but to our knowledge FDA has thus far cleared only locked models. Additionally, to keep the interview-based portion of our empirical inquiry tractable, we focus on institutional adopters rather than individual health care professionals. In the category of adopters, we focus on institutions that provide health care rather than payers. Although scholars have rightly pointed out that payers can play an important role in technology adoption,[20] and we did speak informally with payers,we were unable to identify a critical mass of major payers sufficiently familiar with ML-CD to break them out as a separate group of formal interviewees.

We find that despite developers' desire to keep details regarding training data and training process secret, accountability challenges associated with secrecy have been mitigated by

[20] *See, e.g.*, Rachel Sachs, *Prizing Insurance*, 30 HARVARD JOURNAL OF LAW & TECHNOLOGY 154 (2016); Rebecca S. Eisenberg & W. Nicholson Price, II, *Promoting Health Care Innovation on the Demand Side*, 4 J. L. & Biosciences 3 (2017),

the reality that most ML-CD models that have emerged thus far have involved low to moderate risk. For these models, sufficient accountability can generally be achieved through public disclosure of rigorously collected performance/evaluation data as well as public disclosure of summaries of training data, training process, and resulting model.

That said, our research does suggest that FDA may not always be asking for sufficiently rigorous performance information with respect to the ML-CD software it reviews. Another concern is that, even for ML-CD software that goes through relatively rigorous review, the FDA's public documentation allowing marketing of ML-CD devices often contains almost no information about either the machine learning model used or the training data. We offer suggestions for how FDA regulation could improve information flow without sacrificing innovation incentives.

Our research also indicates that greater patent protection for machine learning software might be useful, particularly for small firms. Although ML-CD patents themselves do no provide much useful information, patents do provide protection that survives more robust disclosure in other venues, most notably peer-reviewed publications. Of course, the potential positive benefits of patents have to be balanced with concerns about ensuring patent quality.

Finally, our research indicates that both developers and adopters are quite concerned about tort liability. This concern notwithstanding, some adopters may accept performance characteristics from "test" data that is merely a specially segregated subset of the training data. To the extent that the adopters are planning subsequently to evaluate, and potentially update, ML-CD models based on their own data, initial reliance on test data may be acceptable. But such reliance absent internal piloting/testing may result in real-world performance that differs significantly from expectations. As we discuss below, tort liability should be structured to

incentivize communication of risk and also to incentivize actors to take appropriate precautions against poor performance.

The Article proceeds in four parts. Part I summarizes intellectual property, FDA, and tort law applicable to ML-CD, with a focus on the disclosure, innovation, and accountability implications of that law. In Part II, we specify our empirical and analytic methodology. Part III then applies this methodology to investigate the current state of information flow and innovation. Part IV provides suggestions for legal and policy improvements that promote accountability but are also sensitive to the need for innovation incentives.

I. THE LEGAL INFRASTRUCTURE: INTELLECTUAL PROPERTY, FDA, TORT LAW

Intellectual property, FDA, and tort law all regulate disclosure, innovation, and accountability in the area of ML-CD software and associated data.[21] Notably, all three legal regimes have struggled to accommodate the relatively recent convergence of software and biomedicine. ML-CD software arguably takes the challenge posed by software in biomedicine to yet another level. Accordingly, the legal frameworks discussed below are far from settled. Although this uncertainty creates obvious challenges for business planning, it also creates opportunities for the types of pragmatic improvements in law and policy that we suggest in Part IV.

We begin with intellectual property law.

[21] Privacy law and legal requirements surrounding reimbursement also affect disclosure and innovation in ML-CD software. Although we do not focus on privacy and reimbursement, we note these considerations where directly relevant.

A. Intellectual Property Law: Disclosure and Innovation Incentives

For ML-CD software, relevant intellectual property principally comprises patents and trade secrecy. As noted, patent and trade secrecy regimes in software and medical diagnostics have shifted considerably in recent years. Because ML-CD software sits at the intersection of software and medical diagnostics, this section reviews the literature on patents and trade secrecy in both software and medical diagnostics. It then discusses recent legal changes and available evidence on impact.

In general, patent and trade secrecy law provide alternative mechanisms for securing competitive advantage from inventions that are expensive to create and develop but inexpensive to copy. Boilerplate patent law requires that the patent document contain disclosure sufficient to allow scientists and engineers of "ordinary skill" in the relevant technology area to "make and use" their invention.[22] Relatedly, applicants are supposed to provide a "written description" that provides insight into the structure of the invention they are claiming.[23] Under standard accounts of the patent system, this disclosure, which mirrors the scientific research and publication norm of reproducibility, is the *quid pro quo* that inventors provide to society in exchange for a time-limited right to control both competition and cumulative innovation in their area of invention.[24]

Meanwhile, trade secret law provides protection only if the information can actually be kept secret. Relatedly, unlike patent law, trade secret law allows a competitor to use freely information it discovers through reverse engineering the competitor's marketed product. For this reason, conventional wisdom holds that firms generally patent information that would otherwise

[22] 35 U.S.C. § 112.
[23] Ariad v. Eli Lilly, 598 F.3d 1336 (2010).
[24] *See* Jacob S. Sherkow, *Patent Law's Reproducibility Paradox*, 66 DUKE L.J. 845 (2017) (noting this classic bargain but also arguing that courts should consider evidence that emerges after patent filing to determine whether the bargain has been met).

be susceptible to reverse engineering.

This conventional wisdom notwithstanding, patent and trade secrecy regimes do overlap somewhat. Two reasons for such overlap are built into the current patent statute. First, the statute allows patentees to impose on those who wish to use the patent some burden of experimentation, so long as the experimentation is not "undue."[25] Some limitation on disclosure may be justified on the grounds that the law wants to encourage potential users/copyists to engage in licensing, with access to the secret information given under the terms of the licensing agreement. On this view, disclosure that allows immediate replication may encourage unlicensed use/copying and place the burden on the patent owner to sue.[26] Second, the statute not only encourages filing of applications relatively early in the R&D process[27] but it doesn't require – or even allow – updating disclosure as more information is gained through further R&D.[28]

Implementation constraints can also create overlap. Most notably, patent examiners operate under onerous time constraints (approximately 19 hours per application) that make it difficult for them to push back against sophisticated patent applicants that desire to maintain patent and secrecy protection over the same information.[29]

Nonetheless, unlike trade secrecy, patent law creates some requirement of affirmative disclosure. Indeed, in some surveys, firms have reported choosing not to patent precisely

[25] 35 U.S.C. § 112.
[26] *See* Arti K. Rai, *Risk Regulation and Innovation*, 92 NOTRE DAME L.REV. 1641, 1647 (2017) (discussing this possibility).
[27] For a critique of this feature of patent law, *see* Christopher A. Cotropia, *The Folly of Early Filing in Patent Law*, 61 HASTINGS L.J. 65 (2009).
[28] For a proposal that such updating should be required, *see* Jeanne Fromer, *Dynamic Patent Disclosure*, 69 VAND. L. REV. 1715 (2016).
[29] Michael Frakes and Melissa Wasserman, *Is the Time Allocated to Patent Examiners Inducing Examiners to Grant Invalid Patents*, 99(3) THE REVIEW OF ECONOMICS AND STATISTICS 550 (2017).

because of this disclosure function.[30] Small firms focused in the life sciences have cited reluctance to disclose information as the most important reason not to patent.[31]

For those firms that do patent, particularly in the life sciences, some empirical literature indicates that the level of affirmative disclosure is sufficient that researchers read the patents[32] and follow-on innovators license them.[33] Even in software, where researcher review of patents is less common, some researchers do turn to patents as a source of information.[34]

In the context of traditional rules-based software, USPTO guidance requires that the patent disclose information about the algorithm(s) and the computer environment used.[35] The USPTO is currently investigating what, if any, additional disclosure it may require going forward from those who file machine learning applications. The agency has issued a Request for Comments (RFC) asking for input on how patent applications that claim models based on machine learning can meet disclosure requirements.[36] This RFC emphasizes disclosure challenges associated with model complexity and unpredictability, particularly for deep learning systems, and asks how meet these challenges.

Notably, enforceable patents may make developers more comfortable with greater disclosure *outside* the patents. Disclosure of this additional knowledge can occur in connection

---

[30] See Stuart Graham et al., *High Technology Entrepreneurs and the Patent System: Results of the 2008 Berkeley Patent Survey*, 24 BERKELEY TECH. L.J. 255, 314 (2009).
[31] *See id.*
[32] Lisa Larrimore Ouellette, *Who Reads Patents*, 35 NATURE BIOTECHNOLOGY 421 (2017).
[33] Deepak Hegde and Hong Luo, *Patent Publication and the Market for Ideas*, 64 MANAGEMENT SCIENCE 652 (2018).
[34] *See* Ouellette, *supra* note __.
[35] Vasudevan Software Inc. v. Microstrategy Inc., 783 F.3d 671, 682-83 (Fed. Cir. 2015); USPTO, *Examining Computer-Implemented Functional Claim Limitations for Compliance With 35 U.S.C. 112*, 84 FR 57 (2019). .
[36] USPTO, *Request for Comments on Patenting Artificial Intelligence Inventions*, 84 FR 44889 (2019). For responses to these comments, *see* https://www.uspto.gov/initiatives/artificial-intelligence/notices-artificial-intelligence

with a patent license (as noted above) or through subsequent publication. The latter phenomenon, which one scholar has dubbed "peripheral disclosure,"[37] is particularly significant because it represents general public disclosure, not simply disclosure to those entities that may license the patent.

Peripheral disclosure is not a trivial phenomenon. To the contrary, as various empirical studies that rely on "patent-paper" pairs demonstrate, the phenomenon is pervasive[38], particularly in the life sciences. For life science academics, even academics who patent, publication remains the most important currency. And for industry developers, peer-reviewed evidence can promote market adoption.

How do patents fare as incentives for innovation? In general, the answer depends substantially on the size of the entity doing the innovation and on technological category. As to size, small firms and startups typically rely more heavily on patents than large firms: while large firms may be able to rely on tools other than patents to attracting the financing necessary to innovate and to maintain competitive advantage from that innovation, small firms are more likely to use patents.[39]

As to technological category, the available evidence indicates that patents provide the strongest incentives in the areas of small molecule therapeutics and traditional medical devices (e.g. standard surgical and dental equipment).[40] In those areas, regulatory requirements are high,

---

[37] Jason Rantanen, *Peripheral Disclosure*, 74 UNIVERSITY OF PITTSBURGH LAW REVIEW 1 (2012).
[38] See, e.g., Tom Magerman et al., *Does Involvement in Patenting Jeopardize One's Academic Footprint? An Analysis of Patent-Paper Pairs in Biotechnology*, 44 RESEARCH POLICY 1702 (2015).
[39] *See, e.g.*, A. Conti, Jerry Thursby, and Marie Thursby, *Patents as Signals for Startup Financing*, 61 JOURNAL OF INDUSTRIAL ECONOMICS 592 (2013); Joan Farre-Mensa, Deepak Hegde, and Alexander Ljungqvist, *The Bright Side of Patents*, NBER Working Paper 21959 (2015).
[40] *See* Bronwyn H. Hall and Dietmar Harhoff, *Recent Research on the Economics of Patents* 4 ANNU. REV. ECON. 541 (2012) (reviewing the empirical literature and noting that positive innovation effects of patents are likely "to be centered in the pharmaceutical, biotechnology, and medical instrument areas . . .")

R&D cycles are lengthy and expensive, and reverse engineering of marketed products is possible.

In contrast, certain medical diagnostic tests, particularly so-called laboratory developed tests (LDTs) run by labs certified by the Center for Medicare and Medicaid Services, have generally not been the subject of FDA regulation. Although the FDA has said it has authority to regulate, and has even proposed guidance on such regulation, thus far it has exercised enforcement discretion.[41] Commentators have argued that the absence of regulatory hurdles for LDTs makes diagnostic innovation comparatively less expensive, and thus a comparatively poorer fit for patenting, than other types of medical innovation.[42]

Even more than diagnostics, software has generally not been regulated and has had low development costs. Firms, including small firms that rely on VC funding, have preferentially relied on first-mover advantages, secrecy, and copyright to capture competitive advantage.[43] Software patents have also been controversial because of concerns over quality (e.g. obviousness, vagueness, and excessive breadth, particularly given dearth of disclosure).[44]

For both diagnostic and software patents, the force of criticism appears to have moved the Supreme Court. Under the test for patent eligibility created by the Court in its 2012 *Mayo v. Prometheus* decision, claims directed to a "law of nature," such as a diagnostic correlation between biomarker and phenotype, are not eligible for patenting unless they cover "something more" than the law of nature itself. The Court's 2014 decision in *CLS Bank v. Alice* similarly

---

[41] *See* Gail Javitt, *FDA Regulation of Laboratory Developed Tests: A Long Saga*, Law 360, December 15, 2016.
[42] Robert Cook-Deegan, Shubha Chandrasekharan, and Misha Angrist, *Against Diagnostic Monopolies* 458 NATURE 405 (2009).
[43] *See, e.g.*, Stuart Graham et al., *High Technology Entrepreneurs and the Patent System: Results of the 2008 Berkeley Patent Survey*, 24 BTLJ 1255, 1290 (2009).
[44] See Arti K. Rai, *Improving (Software) Patent Quality Through the Administrative Process*, 51 HOUSTON L.REV. 503, 504-505, 509-511 (2013).

held that a claim "directed to" an algorithm is not eligible for a patent unless it covers "something more" than the algorithm itself.

In the wake of *Mayo* and *Alice*, securing medical diagnostic patents has become more challenging. According to one study, in the month after *Mayo* was decided, the PTO rejected 32% of the patent applications for medical diagnostics, up from 7% before.[45] The rate of rejections among medical diagnostic applications has continued to climb.[46]

Similarly, securing machine learning patents has become more challenging. According to one study that uses the USPTO's technology classification for "artificial intelligence,"[47] eligibility rejections in that classification have increased since the Supreme Court cases, and particularly since a 2016 lower court decision, *Electric Power Group, LLC v. Alstom S.A.*,[48] further interpreting the Supreme Court cases.[49]

We are not aware of a study that specifically examines rejection rates for machine learning patents that focus on clinical decision making. Nonetheless, one of the implications of the more general trends may be relatively fewer patent ML-CD applications and hence less information flow, either through the mechanism of the patent document itself or – perhaps more

---

[45] Colleen Chien et al., *Decoding Patentable Subject Matter*, 2018 Patently-O Pat. L.J. 1, 15 (Oct. 21, 2018), https://bit.ly/2oBO1i5

[46] *See id.* *See also* Mateo Aboy et al., *Mayo's Impact on Patent Applications Related to Biotechnology, Diagnostics, and Personalized Medicine*, 37 NATURE BIOTECHNOLOGY 513 (2019) (using a different definition of medical diagnostic but also finding increasing rejection rates after *Mayo*).

[47] Kate Gaudry and Samuel Hayim, Artificial Intelligence Technologies Face Heavy Scrutiny at the USTPO, available at https://www.ipwatchdog.com/2018/11/28/artificial-intelligence-technologies-facing-heavy-scrutiny-uspto/id=103762/

[48] 830 F.3d 1350 (Fed. Cir. 2016).

[49] However, recent guidance from the USPTO appears to have reduced rejection rates. Kate Gaudry and Samuel Hayim, *Update on 101 Rejections at the USPTO: Prospects for Computer-Related Applications Continue to Improve Post- Guidance*, available at https://www.ipwatchdog.com/2019/08/13/update-101-rejections-uspto-prospects-computer-related-applications-continue-improve-post-guidance/id=112132/

importantly – through "peripheral disclosure."

What does the reduced availability of patents mean for innovation? Given the disproportionate reliance on patents by small firms, researchers have focused on the VCs that often finance these firms. Survey evidence indicates that VCs focused on software view *Mayo* and *Alice* as less detrimental to their investment plans than do VCs focused on medical devices, biotechnology, and pharmaceuticals.[50] However, prior to our study, the literature had not specifically examined VC response in the area of ML-CD. Because ML-CD represents neither a a conventional medical device nor conventional software, but instead the type of intersectional technology that is likely to represent the future of a general purpose technology like machine learning, the inquiry is particularly important. (Our study's findings are discussed in Part III.)

The next section introduces the role that FDA law plays along the dimensions of information flow, incentives for innovation, and accountability.

B. FDA Law: Information Flow, Innovation Incentives, and Accountability

Like intellectual property law, law specific to the FDA has an impact on information flow and innovation. The impact on information flow and innovation follows from FDA's primary mission of promoting accountability through risk-benefit analysis. For those products over which it exercises jurisdiction, the FDA may facilitate information flow not only by reviewing the information itself but also by making certain data publicly available. As for innovation, regulatory requirements for safety and efficacy creates obvious barriers for firms seeking to achieve market entry for their products. However, for those firms that prove their ability to innovate along the dimensions of safety and efficacy, these requirements serve as an entry barrier

[50] David Taylor, *Patent Eligibility and Investment*, 41 CARDOZO LAW REVIEW __ (2019).

against competitors and thereby allow innovator firms to recoup R&D costs.[51]

In general, the Food, Drug, and Cosmetic Act (FDCA) that FDA administers confers on the agency broad power to regulate as medical devices non-chemical items that are used for disease diagnosis, mitigation, treatment, or prevention.[52] The Cures Act specifically addresses software, indicating that software can be subject to FDA regulation unless it falls within a set of specific exemptions. For example, "administrative support software" – defined to include not only standard billing and workflow software but also software that performs "population health management" – is exempt from regulation.[53]

Notably, both developers and the FDA appear to have included within the exempted category of "population health management" tools that forecast the need for supplemental support for specific patients.[54] Thus the determination of what counts as "population health management" may be quite relevant to ML-CD accountability.

The Cures Act also exempts from regulation certain software that "support[s] or provid[es] recommendations to a health care professional about prevention, diagnosis, or treatment of a disease or condition."[55] In order to qualify for this exemption, the support software must meet two conditions. The first condition relates to inputs. Support software is not exempt if it receives as input a medical image, "a pattern or signal from a signal acquisition system," or data from an *in vitro* diagnostic device.[56] The second condition addresses opacity:

---

[51] For a perceptive discussion of the FDA's role in information flow and innovation, see Rebecca S. Eisenberg, The Rorle of the FDA in Innovation Policy, 13 MICH. TELELCOMM. & TECH. L. REV. 345 (2007).
[52] 21 USC 321(h))
[53] FDCA § 520(o)(1)(A).
[54] *See, e.g.,* Obermeyer, *supra* note __ (algorithm that determines which patients are deemed high-risk and therefore receive additional intervention).
[55] FDCA § 520(o)(1)(E)(3)
[56] FDCA § 520(o)(1)(E)(1).

even if the software *doesn't* use image, signal, or *in vitro* diagnostic data, it falls under FDA authority unless it allows the health care professional to "independently review" the basis for its recommendation.[57]

Interpreting the "support" exemption of the Cures Act has proved to be challenging. One challenge has involved determining how the Act's use of the term support relates to a tripartite categorization for clinical decision making software the FDA had earlier developed as part of the International Medical Device Regulators Framework (IMDRF). The latter categorization does not use the term "support" – instead, software may "inform" clinical management, "drive" clinical management, or "treat/diagnose."

The FDA's latest guidance on the Cures Act, issued in September 2019,[58] invokes the tripartite IMDRF categorization in its determination that "support" software encompasses only software that "informs" clinical management. According to the FDA, the two other IMDRF categories always fall within its regulatory jurisdiction.[59]

Interpreting the Act's "independent review" criterion for exemption of support software has also been challenging. The FDA's 2019 guidance states that in order for support software to satisfy the independent review criterion, the software developer needs to provide the software intended use/user; "the inputs used to generate the recommendation" and the "basis for rendering a recommendation."[60] Providing the basis for recommendation includes describing the underlying data used to develop the algorithm.[61] It also includes providing specific supporting

[57] FDCA § 520(o)(1)(E)(4).
[58] FDA, CLINICAL DECISION SUPPORT SOFTWARE, DRAFT GUIDANCE FOR INDUSTRY AND FDA STAFF (2019) available at https://www.fda.gov/media/109618/download.
[59] *Id.* at 13-14.
[60] *Id.* at 12.
[61] *See id.*

sources for the recommendations, such as clinical practice guidelines or published literature. Indeed, all of the examples of independent review in the guidance appear to involve rules-based software that uses clinical guidelines, FDA labeling, or "generally accepted clinical practice."[62]

One ambiguity raised by the guidance's approach to independent review is whether a ML-CD model that can provide a generally accepted rationale for specific recommendations at the point of use (for example, because another type of machine learning running on top of the ML-CD "explains" key factors), but is nonetheless relatively opaque as to its overall model, would be exempt from review. Additionally, as we discuss below, estimates of the certainty of the recommendation provided by the model are important for independent review. Yet the guidance does not speak to the issue of certainty.[63]

FDA regulation can have a powerful impact on information flow. For example, the FDA has the power to require submission of trade secret information and, indeed, regularly does so in the context of clinical trial data on biopharmaceuticals and manufacturing data for biologics. Under the Freedom of Information Act (FOIA),[64] however, it is permitted to withhold this information from the public under the broad category of "confidential commercial information." Given the FOIA exemptions and laws like the Trade Secrets Act that subject government employees who disclose trade secrets to criminal sanctions, developers should, in principle, be comfortable providing FDA details about model training and data.[65]

---

[62] *Id.* at 17-18.

[63] The challenge of parsing of what constitutes independent review is mitigated by FDA's consistent focus on risk. Specifically, in cases where software would not be exempt because it could not be independently reviewed, the agency says it will take a risk-based approach. *Id.* at 17. In general, this means that the agency will not exercise oversight if the support software is used in connection with medical conditions that are not serious. Id. The upshot is that the independent review criterion is unlikely to loom very large for software that addresses non-serious cases.

[64] 5 U.S.C. § 552

[65] Indeed, disclosure to the FDA could even extend to the level necessary to yield at least some level of

In principle, then, the exercise of FDA jurisdiction could improve accountability even in cases where FDA cannot publicly reveal the information over which it has custody.

C. Tort Liability: Accountability, Disclosure, Innovation

Like FDA law, tort law takes as one of its goals appropriate balancing of risk and benefit. Tort liability may also effects on disclosure. Here we focus on one area ML-CD tort liability: product liability law and how it may shape design and disclosure decisions made by developers. We do not address potential negligence liability faced by physicians who implement (or don't implement) the recommendations made by ML-CD.[66]

In general, the law surrounding product liability for software *qua* software (as contrasted with software embedded in a product) is most notable for a dearth of successful claims.[67] For plaintiffs, the ability to bring successful claims has been complicated by a number of doctrinal disputes. These include: uncertainty over whether software qualifies as the type of product traditionally subject to the torts framework or is instead an intangible informational service; the historical absence of physical injury associated with software malfunction (as contrasted with "economic loss," for which the tort system does not compensate); and challenges in proving that specific software defects caused the injury.[68]

As commentators have persuasively argued, however, it is not clear that software *per se*,

---

literal reproducibility – for example, all details regarding the development process, such as training data, learning algorithm, tuning process, parameters, and hyper-parameters. That said, given its current resources, the FDA's ability to use information relevant to reproduction profitably is unclear.

[66] Commentators have written very perceptively about how the custom-based standard of care for medical malpractice liability is likely to affect the actions of individual physicians who use ML-CD software. *See, e.g.*, W. Nicholson Price et al., *Potential Liability for Physicians Using Artificial Intelligence*, 322 JAMA 1765 (2019). For a more general discussion of negligence issues that can arise when humans interact with machine learning to make decisions, *see* Andrew D. Selbst, *Negligence and AI's Human Users*, 100 BOSTON UNIVERSITY LAW REVIEW __ (2020).

[67] For a review of the extensive literature noting the absence of successful claims, *see* Bryan Choi, *Crashworthy Code*, 94 WASHINGTON LAW REVIEW 39 (2019).

[68] *Id.* at 62-79.

including machine learning software, should be treated differently from software embedded in a device. As they have noted, lawsuits against faulty software embedded in a device (e.g. faulty software that caused inexplicable acceleration in cars manufactured by Toyota) have proceeded without doctrinal hiccups. [69]Assuming doctrinal challenges associated with product liability for software could be overcome, what might such liability look like in the ML-CD context? As a threshold matter, product liability (as contrasted with medical malpractice liability) would most likely be operative if the software was playing a significant role in clinical decision making – that is, doing more than informing the physician of a decision she could independently review. In that case, the software would have gone through the FDA process, and the state tort liability preemption regime developed by the Supreme Court for that circumstance would presumably apply. Accordingly, we focus here on software that has gone through FDA review.

The Supreme Court has held that for those devices that go through the most rigorous level of FDA review – premarket approval (PMA) – such approval typically preempts liability in state tort law for injuries caused by the products.[70] The PMA track requires submission of technical information as well as evidence regarding non-clinical and clinical investigations. If the ML-CD software goes through a less rigorous level of review, such as 510(k) clearance or *de novo* review, liability might be possible.[71]

Of the various theories of product liability, design defect and failure-to-warn liability

[69] *See* Bryan Casey, *Robot Ipsa Loquitur*, 108 GEORGETOWN LAW JOURNAL 225 (2020) (making this argument using the example of In re Toyota Motor Corp. Unintended Acceleration Mktg., Sales Practices, & Prod. Liab. Litig., 978 F. Supp. 2d 1053, 1100-01 (C.D. Cal. 2013) and further arguing for address questions of causation through the doctrine of *res ipsa loquitur*).
[70] Riegel v. Medtronic, 552 U.S. 312, 321-323 (2008).
[71] For products that go through 510(k) review, liability preemption clearly does not apply. *See* Medtronic v. Lohr, 518 U.S. 470 (1996). We are not aware of cases that have assessed preemption with respect to so-called *de novo* review, a level of scrutiny less rigorous than PMA approval but more rigorous than 510(k) clearance.

may have the most significance for ML-CD.[72] Generally speaking, a design defect may be found if the product's foreseeable risk of harm "could have been reduced or avoided" by the developer's adoption of a "reasonable alternative design."[73] In principle, for example, a developer's failure to collect appropriate training data might be seen as a design defect.

As for failure to warn claims, which follow from the premise that a manufacturer has a duty to provide adequate warning of danger associated with product use and instruction on how to use its product safely, the learned intermediary doctrine that applies in health care context makes health care professionals the appropriate recipient of the warning.[74] In principle, if a developer failed to warn hospital or physician adopters of the limits of its model, it might be held liable. Appropriate warnings might include statements that the model should only be used with certain types of input data or statements regarding the level of certainty associated with specific recommendations.

This shifting legal landscape provides the context for our quantitative and qualitative inquiry into information flow and innovation incentives in ML-CD. We next describe our methods.

II. METHODS

To investigate innovation incentives and information flow empirically, we employed both quantitative and qualitative methods. On the incentives side, we were particularly interested in what impacts, if any, shifts from patenting to secrecy were having on the small firms that tend to

[72] Manufacturing defect liability, which can be found if a mass-produced product differs "from its siblings in a manner that makes it more dangerous than the others," *see* Casey v. Toyota Eng'g & Mfg. N. Am., Inc., 770 F.3d 322, 329 (5th Cir. 2014) is possible but seems less likely in the context of ML-CD models, particularly ML-CD models that are not mass produced but are cloud-based.

[73] RESTATEMENT (THIRD) OF TORTS: PRODS. LIAB. § 2(b) (AM. LAW. INST. 1998). Some states

[74] *See* RESTATEMENT (THIRD) OF TORTS: PROD. LIAB. § 6 cmts. d-e (1998).

rely disproportionately on patents. Accordingly, we gathered and analyzed time series data from Pitchbook® on VC investment in ML-CD software.

The Pitchbook® platform calculates individual deal-level venture capital and private equity investment through analysis of regulatory filings, news releases, and websites.[75] One category into which Pitchbook places investment deals is "artificial intelligence and machine learning" (AI/ML). Pitchbook also categorizes deals by industry. Our search, conducted on December 15, 2019, intersected the AI/ML category with "decision/risk analysis through health care technology systems," the Pitchbook-defined industry that appears most relevant to clinical decision software. We then reviewed publicly available materials on the 55 firms retrieved by the search results. All of these firms operated in the health space and used terms like "AI" or "machine learning" in their marketing materials. This review gave us confidence that our search did not retrieve a large number of false positives.[76] However, both Pitchbook coverage and our search may be underinclusive.

We also investigated both issues surrounding disclosure and innovation incentives through the use of qualitative methods. We held a private workshop to which we invited large, small, and academic developers, institutional adopters, insurance firms, patent lawyers, and FDA officials. The invitee list for the private workshop was generated through a combination of convenience sampling and snowball sampling. Discussion at this workshop helped us generate hypotheses, identify individuals for formal interviews, and frame questions for these formal interviews.[77] In general, representatives from VC firms are often particularly difficult to secure for formal interviews. Through this private workshop, and through a Pitchbook search of VC

[75] https://pitchbook.com/research-process
[76] Because of the overclaiming associated with both marketing materials and use of terms like "AI" and "machine learning," we cannot be confident there were no false positives.
[77] Topics covered in the interviews are listed in Appendix A.

firms most active in the overall category of health AI/ML, we were also able to identify VCs for formal interviews.

We did not use statements made in the private workshop as the foundation for any of our inferences. However, where the inferences we drew from formal interviews were supported by statements made in the private workshop, we note that point below..a

As shown in Table 1, we conducted formal interviews lasting between 30-45 minutes with the following categories of relevant stakeholders: 1) venture capitalists that invest in early and later stage ML- CD startups; 2) small developers (including developers based in academic institutions); 3) large developers; and 4) institutional adopters/providers. The institutional adopters/providers included both research-oriented academic medical centers and others types of relatively large providers.

Pursuant to the terms of our IRB protocol, we informed all of our interviewees that any information they provided us would not be used to identify them. Accordingly, our discussion refers to interviewees by occupational category only.

Table 1 sets out the distribution of our formal interviewees. The specific questions we asked interviewees differed by category of interviewee; we list topics covered for each category of interviewee in Appendix A.

Table 1

Category	**Firms Interviewed**	**Individuals Interviewed**
Venture capital	4	5
Small and academic developers	5	6
Large developers	3	8
Institutional adopters/providers	5	7

Finally, we identified publicly available information on four ML-CD devices for which FDA has granted a *de novo* request. We focused on *de novo* review because, to our knowledge, this is the most rigorous review that FDA has used thus far for ML-CD devices. The FDA disclosure associated with this review is more extensive than for the ML-CD devices that have used the 510(k) pathway.[78] We were, however, unable to find actual labeling information in the FDA disclosure. Instead, the disclosure provided summary information.[79]

Additionally, all four of these devices have patents and/or peer-reviewed publications that provided additional disclosure. Accordingly, the public disclosure associated with these devices likely represent something of a "high water mark." To the extent that public disclosure associated with these devices falls short, there is reason to believe that disclosure problems may be even greater for other ML-CD products.

III. RESULTS: INNOVATION INCENTIVES AND DISCLOSURE

A. Quantitative Results on Incentives

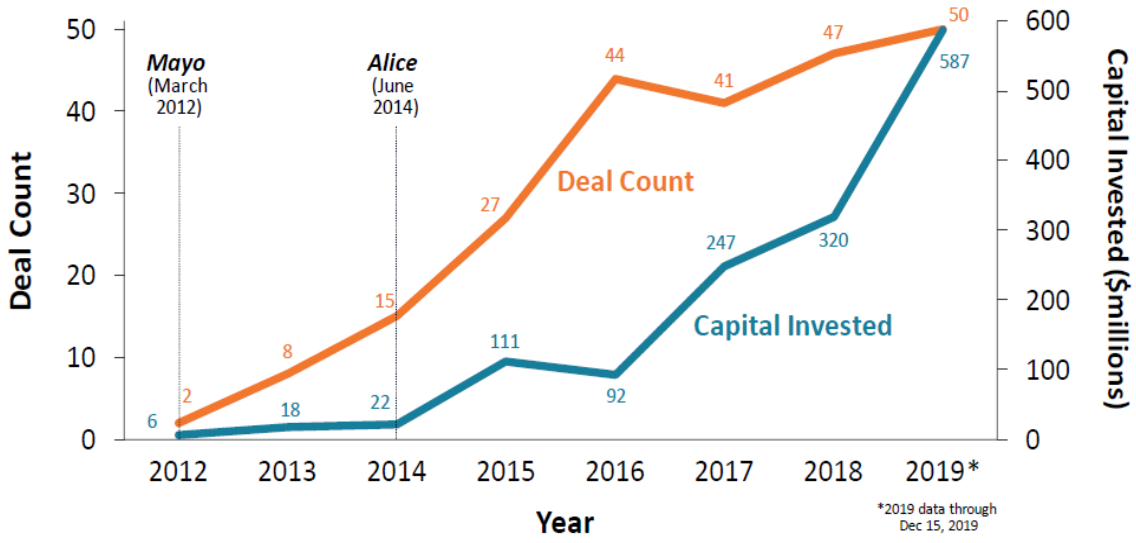
Figure 1 presents data from January 1, 2012 (the year of the *Mayo* decision and also the first year Pitchbook records any significant VC/PE activity in the area of ML-CD) through December 15, 2019. As the data show, uncertainties regarding patents have not thwarted a 26-

[78] Notably, these 510(k) clearances may use as their predicate device inventions that do *not* operate via machine learning. For example, the 510(k) FDA clearance for the Arterys CardioDL device, see https://www.accessdata.fda.gov/cdrh_docs/pdf16/K163253.pdf, doesn't mention machine learning. Moreover, the predicate device (itself the result of a chain of predicate device references) doesn't appear to utilize machine learning. In contrast the advertising for Cardio DL invokes the device's "trained deep learning algorithm." *See* https://www.arterys.com/press-release/arterys-receives-fda-clearance-for-the-first-zero-footprint-medical-imaging-analytics-cloud-software-with-deep-learning-for-cardiac-mri/

[79] Because we wanted to assess information that was readily available in the public domain, we did not submit Freedom of Information Act requests for the full submission dossiers. We were also told informally that any response to such a request would be heavily redacted.

fold increase in VC/PE investment from 2014-2019.

Figure 1: AI/ML investment in decision/risk analysis through health technology



Of course this data does not speak to what investment would have looked like absent patent-related challenges. To investigate that question, one comparator that we examined was AI/ML investment in biotechnology and pharmaceutical development.[80] In the case of biopharmaceutical development, the AI/ML is only a research input, either for discovery or for streamlining the development process. Unlike ML-CD models, the ultimate drug or biologic was patent-eligible throughout the time period we examined. The secure availability of patent protection at the end of R&D, and the concomitant secure ability to price at supra-competitive levels, arguably buffers the biopharmaceutical R&D process from the effects of *Mayo* and *Alice*.

Here the investment data from January 1, 2012 through December 2018 (Figure 2) do

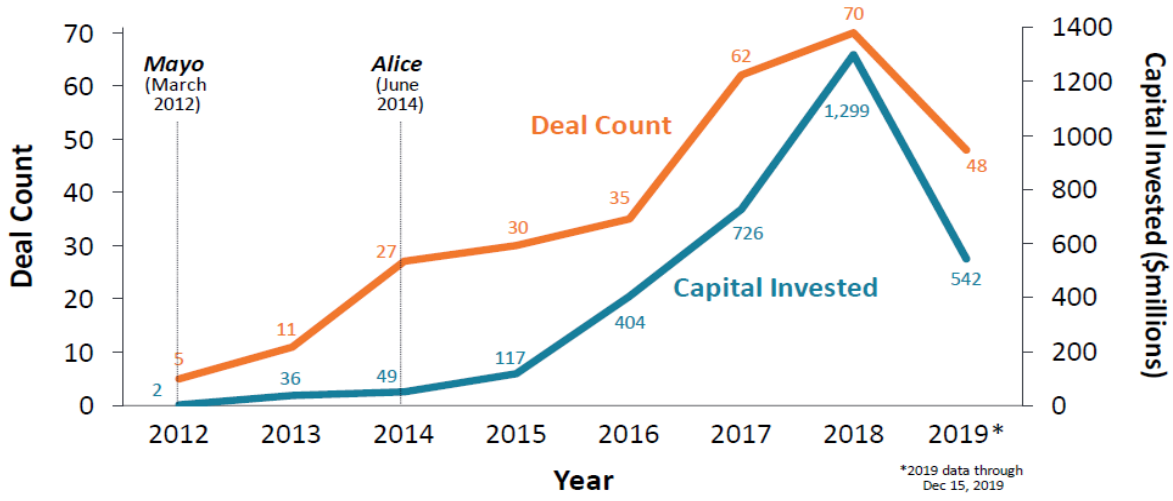
[80] More specifically, we intersected the Pitchbook AI/ML "vertical" category with the pharmaceuticals and biotechnology "horizontal" category.

show a greater increase in VC/PE interest in biopharmaceuticals through 2018. (There was an unexplained decline in 2019). While VC/PE capital invested in the use of AI/ML for biopharmaceutical R&D increased 26-fold from 2014-2018, VC/PE capital invested in its use for decision/risk analysis during 2014-2018 increased 14-fold.

VC/PE investment in biopharmaceutical AI/ML has limits as a comparator. As an initial matter, although overall profits from biopharmaceutical R&D may be buffered from the effects of the Supreme Court decisions, small AI/ML firms relying on secrecy rather than patents in their licensing negotiations with larger firms might be in a weaker bargaining position and thus obtain a smaller share of the overall profit distribution. The consequence might be diminished VC/PE enthusiasm for these firms.

Additionally, for biopharmaceuticals, regulatory and insurance schemes were relatively stable during the period studied. In contrast, regulation of, and insurance for, ML-CD models was in flux. Accordingly, we do not purport to draw causal inferences from the fact that growth in VC/PE capital investment was, during the period 2014-2018, higher for AI/ML used in biopharmaceutical R&D than for AI/ML used in decision/risk analysis. Instead, we view our comparative results as indicating that while patents are hardly a *sine qua non* for VC-backed ML-CD products, they may be helpful.

Figure 2: AI/ML investment in biopharmaceutical R&D



We turn next to our qualitative investigation of the range of issues surrounding innovation incentives and disclosure.

B. Interviews on Innovation Incentives and Disclosure

All of the VCs and developers (small and large) that we interviewed pointed to training data as a key component of competitive advantage. Most thought it was more important than any other competitive asset.[81] For this reason, although VCs and developers were willing to share summary statistics (including demographic and other subgroup statistics) with third parties, they were not willing to share actual data.[82] For all but one of these VCs and developers, third parties

[81] Participants in the private workshop also noted the competitive importance of training data and noted that it was not being shared with the FDA.

[82] Sharing data that has not been de-identified could also raise privacy issues. However, our interviewees focused on competitive advantage.

with whom they were not interested in sharing included the FDA (assuming of course that the FDA did not condition reaching market on such a requirement).

One academic developer did note, however, that developers *should* be willing to share training data with the FDA, as the agency is capable of maintaining such data as a trade secret. According to this developer, summaries of training data may be insufficiently informative, particularly when the model's performance across different settings is based in part on subtle artifacts of the original training data's institutional setting that may not be captured through summary statistics.[83]

All developer interviewees agreed that thus far they were not aware of cases where the FDA has asked for full training data. Several noted constraints on FDA's institutional capacity. Some interviewees also noted that FDA institutional capacity constraints were one reason why adopters should begin to increase their own expertise.

In addition to data, all VCs and developers who spoke to the issue viewed details regarding the trained model as a key competitive asset. For example, in the context of neural networks, one small developer stated that details regarding network weights and numbers of layers and nodes were an important asset to keep confidential, even with respect to the FDA. Indeed, the published literature shows that even large developers like Google's Deep Mind may release basic learning algorithms but keep improvements that lead to the trained model secret.[84]

On the question of patents, VC and developer interviewees had heard that machine learning patents were difficult to secure and enforce. One large developer noted that it viewed

[83] Zech, J. et al., Confounding variables can degrade generalization performance of radiological deep learning models, available at https://arxiv.org/abs/1807.00431

[84] Tomasev, N. et al., A clinically applicable approach to continuous prediction of future acute kidney injury. 572: 116-119 (2019) (noting that team's "experimental framework makes use of proprietary libraries")

patents as useful primarily for defensive reasons (that is, to neutralize threats from competitor firms that wanted to sue on their patents).

In contrast, one VCs and one small developer did think that if patents could be made easier to secure and enforce, they could be valuable for purposes beyond defensive use. One VC conveyed his strong opinion that, in order to incentivize investment in small machine learning firms, the Supreme Court or Congress should revise patent eligibility jurisprudence. This VC stressed the role patents can play not only in attracting capital but also in providing the protection against copying necessary to allow information flow unconstrained by nondisclosure agreements. This VC specifically mentioned publication in peer- reviewed journals.

Some developers also expressed their views on what institutional adopters wanted. One large developer volunteered an opinion that, to the extent that adopters wanted information about the machine learning software, they were generally interested in performance data and, at most, summary information about training data. Only a "small minority" were interested in model details.

All of the institutional adopters we interviewed agreed they didn't necessarily need to know model details. One adopter noted that even major health systems may have not the human capital necessary to make use of model details. Another noted that, even if protected by a nondisclosure agreement, developers were not likely to part with model details, let alone its source code.

Moreover, unlike commentators that have stressed the important of gleaning performance data by using a locked, trained model on completely independent data sets (that is, data that is not simply a specially segregated set of the training data set),[85] adopters to whom we spoke did

[85] Yun Liu et al., *How To Read Articles that Use Machine Learning: Users' Guides to the Medical Literature*, 322 JAMA 1806, 1808 (2019).

not emphasize independent validation, particularly for lower-risk software and/or if they were planning to independently test the software with their own data. They did, however, express a general interest in performance and also in demographic and other subgroup characteristics of the training population.

C. Case Studies

We supplemented our interview-based qualitative research with examination of publicly available information on four prominent ML-CD devices that use supervised machine learning and on which FDA has granted *de novo* review requests. These devices are comparable in that they all use some form of neural network for image-based diagnosis. Additionally, they are all products created by small startup firms.[86] This comparability facilitates comparison but may limit generalizability.

Like devices approved using the premarket approval (PMA) path, *de novo* devices do not have a predicate. And like PMA, the *de novo* pathway produces a fair amount of publicly available FDA documentation. Since no ML-CD approvals appear to have gone through the PMA pathway thus far, these devices represent something of a "high water mark" for purposes of FDA-regulated public disclosure and innovation in the area of ML-CD.

In chronological order of approval, the four devices are Quantitative Insights' (now QClarity)'s QuantX, which screens for cancerous lesions in breast tissue; Viz.AI's ContaCT, a stroke triage and notification system; IDx's IDx-DR, which screens for diabetic retinopathy; and

[86] These four examples do not necessarily represent the universe of ML-CD devices that have reached market via the *de novo* pathway. FDA does not, to our knowledge, maintain a list of devices it considers to be ML-CD. Similarly, publicly available FDA documents do not always specify whether a device uses machine learning. In two cases involving grants of *de novo* review requests, the Apple electrocardiograph (ECG) software, https://www.accessdata.fda.gov/cdrh_docs/reviews/DEN180044.pdf, and the DreaMed Advisor Pro, https://www.accessdata.fda.gov/cdrh_docs/reviews/DEN170043.pdf, we could not be confident that the device used supervised machine learning. We therefore do not discuss those cases.

Imagen's Osteodetect, which detects distal radius fractures.

For each of these devices, we looked for FDA documentation, patents, and publications. In line with recommendations recently put forward by Liu et al.,[87] we examined this documentation for information in the following categories: 1) training data (e.g. input, output, labeling mechanism, amount of data); 2) training method (e.g. machine learning algorithm used, training process, resulting model); 3) evaluation process (e.g. test data, completely independent data set (retrospective or prospective), performance metrics). (See Table 2)

QuantX

In July 2017, FDA granted a *de novo* review request that allowed marketing of the QuantX device, which is intended to assist radiologists in reading MRIs to detect abnormalities suspicious for breast cancer. The FDA decision summary[88] contains substantial information regarding training data. The summary information includes details of MRI acquisition environment, image inclusion criteria, age, and the process by which the label of benign or malignant was determined. Information about race or ethnicity is not provided.

As for training method, the FDA document notes the use of a combined feature score algorithm "based on literature described in detail within the submission." The decision summary document does not, however, itself disclose this information. Similarly, the patents associated with the QuantX device[89] do not disclose any details regarding either the learning algorithm or the trained model.

In contrast, the published literature describes in some detail the training and independent validation of a Bayesian neural network to assign a probability of malignancy to a lesion. In a

[87] *See* Liu et al., *supra*
[88] https://www.accessdata.fda.gov/cdrh_docs/reviews/DEN170022.pdf
[89] US9208556; US9536505

2011 paper, for example, the University of Chicago team discusses a Bayesian neural network that uses 12 features (themselves identified in prior literature). The filing of a patent application in November 2010 may have made the University of Chicago authors, and their institutional employer, comfortable with disclosing more detail in subsequent publications. The extent to which the 2011 model conforms with the model that is the subject of the FDA application is not entirely clear, however.

With respect to evaluation, part of the performance data appears to come from the same images that constitute the training data. The FDA decision summary highlights this point, stating that this "standalone study . . . used the Similar Case Database (i.e. training database)" and therefore "cannot be considered an independent validation study."

The QuantX developer also conducted an entirely independent multiple reader, multiple case clinical study of 111 previously classified cases in which 19 radiologists participated. The study included a sequential reading design, in which the first read was performed using conventional tools, while the second read used the QuantX interface. The FDA decision summary contains detailed information about how these cases were acquired and classified and regarding the training of the 19 radiologists. On the demographic and other subgroup side, it gives age distribution but no information regarding race or ethnicity.

Viz.AI ContaCT

The second FDA grant of a *de novo* review request, in February 2018, addressed Viz.AI's ContaCT, a device that triages suspected large vessel occlusions.[90] The device is intended to operate in parallel with standard workflow in an emergency room.

Neither the decision summary nor the relevant Viz.AI patent[91] contains information on

[90] https://www.accessdata.fda.gov/cdrh_docs/reviews/DEN170073.pdf
[91] US10346979

training data or training process. The FDA decision summary states that performance was evaluated on 300 CT images. Other than noting that the images came from two different clinical sites, the summary does not contain any information about how the 300 CT images were acquired or regarding their demographic/other subgroup characteristics. Some of the detail regarding the 300 CT images appears to be deemed trade secret information that is redacted under the relevant FOIA exemption.[92]

Since the time Viz.AI has been approved, two different groups of clinicians, neither of which appears to be affiliated with Viz.AI, have published abstracts giving performance data that uses retrospective studies on CT images gathered in their clinical settings.[93] These abstracts state that Viz.AI uses a convolutional neural network.

IDx-DR

FDA's third grant of a *de novo* request, in April 2018, involves software that analyzes retinal images and provides primary care clinicians with a recommendation regarding whether diabetic retinopathy (DR) has been detected such that the patient should be referred to a specialist. The primary care physician is not expected to review these images herself.

The FDA decision summary document, the IDx-DR patent,[94] together with peer-reviewed publications that have issued since IDx filed its patent application, provide a substantial amount of information. Notably, however, none of these documents provide information about the training data.

As for training process, various IDx publications disclose that IDx-DR uses convolutional

[92] Id.
[93] https://jnis.bmj.com/content/10/Suppl_2/A101.2; https://www.ahajournals.org/doi/abs/10.1161/str.50.suppl_1.WMP16?af=R&
[94] US Patent 10,115,194 B2

neural networks to identify a number of "independent, validated detectors" for lesions characteristic of DR. IDx-DR then uses a separate machine learning algorithm to combine the detectors into a disease level output.[95] Another machine learning algorithm, implemented as multiple independent detectors, monitors input quality.

On the evaluation front, IDx-DR secured FDA's grant of its review request through a prospective clinical study with 900 patients who were enrolled at 10 primary care sites. The publicly available documentation on this study specifies the inclusion criteria for participants, their demographic and other subgroup characteristics, the training of the operators who collected retinal images that were processed using IDx-DR as well as the training of the operators who collected subsequent retinal images on the same patients that were then read by experts.[96] The FDA grant of the *de novo* request discusses in great detail key performance statistics as well as key statistics indicating how use of IDx-DR would fit into the workflow of a primary care office.[97]

Osteodetect

The fourth grant of a *de novo* review request, involving Imagen's Osteodetect in May 2018, covers a device that detects distal radial fractures. The FDA decision summary does not say anything about training data or training process.

As for data on model performance, the FDA decision summary discusses a retrospective study of 1000 images from an "independent dataset that had not been used for model development." According to the decision summary, these 1000 images were randomly sampled

[95] Abràmoff, M.D., Lavin, P.T., Birch, M. *et al.* Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. *npj Digital Med* **1,** 39 (2018) doi:10.1038/s41746-018- 0040-6

[96] Id.

[97] https://www.accessdata.fda.gov/cdrh_docs/pdf18/DEN180001.pdf

from a "validation database." It is not entirely clear whether this "validation database" represents a separately segregated subset of the original training data or a completely independent data set. The decision summary discusses how ground truth for the 1000 images was established and notes that model performance did not vary by sex, age, or post-surgical status.

The developers also submitted a retrospective study of 400 cases from the same "validation database" that had been reviewed by 24 clinicians. Each clinician independently read each case, both aided by the OsteoDetect device and unaided.

A September 2018 publication by Imagen scientists[98] describes in some detail training data (34,990 wrist radiographs), training process (deep convolutional neural network with "free parameters"), and evaluation process (two different test sets, one randomly withheld from the training set and the other a consecutive sample) for a distal radius fracture detection device. How the model discussed in this publication corresponds to the subject of the FDA application is not clear, however.

Table 2: Publicly Available Information

Device	Device Function	Patent	Publications	Training Data for *De Novo* Requests	Description of Machine Learning Algorithm, Training Process	Evaluation Process

[98] Lindsey. R. et al. Deep Neural Network Improves Fracture Detection By Clinicians. PNAS 115(4): 11591-11596 (2018).

QuantX	Detection of abnormalities suspicious for breast cancer	Yes	Yes	543 images gathered from different locations	None in FDA document; 2011 publication describes Bayesian neural network	Testing on same cases used for training; entirely separate dataset of 111 images
ContaCT	Triage for images suspicious of stroke	Yes	Abstracts published by independent authors	Discussion not found	According to abstracts published by independent authors, CNN	300 CT angiogram images from 2 clinical studies
IDx-DR	Detection of diabetic retinopathy	Yes	Yes	Discussion not found	CNN that combines "independent, validated detectors" for lesions characteristic of DR	Prospective study of 900 patients enrolled at 10 primary care sites
OsteoDetect	Detection of distal radial fracture	None found	Yes	No discussion in FDA document; unclear if training data specified in 2018 article was used in FDA review	None in FDA document; if same model used as in 2018 article, DCNN with "free parameters"	3 different evaluation studies; 2 in FDA document; one (perhaps using different model) in 2018 article

Our case studies illustrate that, at least for those ML-CD applications that go through the *de novo* pathway, some public information regarding the evaluation process is generally available. In contrast, even summary information about training data and training process can be harder to find.

IV. IMPROVING THE INFORMATION ECOSYSTEM

We conclude by offering suggestions on how to improve the innovation ecosystem. We begin with a simple framework informed by the preceding analytic and empirical discussion. Figure 3 frames the challenge in terms of accountability and reproducibility. In our framing, which aims to preserve innovation incentives,[99] an ideal system would operate in the lower right quadrant: it would offer comprehensive accountability with little reproducibility by competitors.

This ideal system does not exist. Nonetheless, significant accountability can be achieved without allowing full reproducibility. In part, this is because the need for accountability depends on risk – with risk measured (per the IMDRF framework) as a combination of seriousness of condition and importance of software to the decision. For example, where overall risk is low to moderate, the requisite level of accountability could likely be achieved through public disclosure of the information contained in oval B.

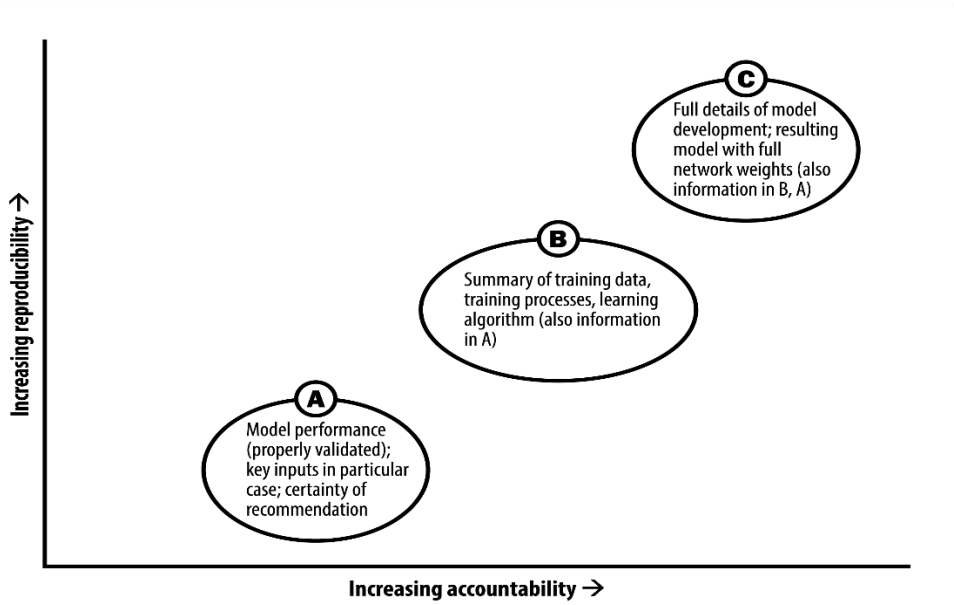
Accountability also depends on the technical capacity of the audience to which the disclosure is directed. For example, if accountability is seen through the lens of an institutional adopter with limited technical capacity to understand full details of model development (full training data set, source code for model), the accountability level achieved by disclosure of the information in oval B could be similar to that achieved by disclosure of information in oval C. And if the software were low to moderate risk, this level of accountability would be sufficient.

In the case of high risk software, particularly software that isn't independently reviewable, this software would almost certainly be regulated by the FDA. In that case, the full

[99] For purposes of our framing, we assume some version of our current intellectual property system, which relies on the ability to exclude, remains in place. We do not consider alternatives such as prizes or comprehensive public funding of all ML-CD research and development.

details of model development could be shared with the FDA. Although such sharing could make the software reproducible (at least in theory, assuming adequate technical capacity on the part of the FDA), the FDA could maintain that information as a trade secret.

Figure 3: Accountability and Reproducibility



How does the empirical reality compare to this framing? Although the current regime does allow for some accountability, there is also room for improvement. The FDA, Congress, institutional adopters, and developers could all play important roles.

As noted, our interviews indicate that the FDA has substantial credibility among ML-CD developers. Thus, the FDA could, without compromising valuable trade secret information, make summary information regarding training data and training process publicly available. In addition to being valuable in itself, such information could help connect the dots as to whether processes and models discussed in academic and trade publications are similar to (or the same

as) those reviewed by the FDA.

Additionally, the agency should generally ask for evaluation on data collected from a source entirely separate from the training data set. At a minimum, public disclosures from the agency should be clear about whether the company used completely independent evaluation data.

In the highest risk cases, the FDA might also ask for access to full training data and source code for the trained model. As the FDA has itself noted, "(r)obust algorithms typically require the availability of large, high-quality, and well-labeled training data sets."[100] This data would clearly be protected from full public disclosure.

To be sure, as many of our interviewees noted, the FDA's institutional capacity for full reproducibility review is limited. Congress might therefore consider allocating resources for development of such institutional capacity. That said, we recognize that allocating resources for one goals often means diverting resources from other goals, which can represent a substantive and political economy challenge.

Finally, our quantitative and qualitative results also suggest a need for sensitivity to the role of innovation and disclosure incentives, particularly for small firmsStronger patent protection for ML-CD products is not necessarily indicated, particularly given the concerns about poor quality that motivated *Mayo* and *Alice*. However, such patents could help small firms attract investment. Moreover, patents provide protection that survives disclosure, including in the type of peer-reviewed publications that are (as they should be) important for achieving market adoption.

As Congress evaluates how to respond to the Supreme Court's patent decisions, it should

[100] FDA, DEVELOPING A SOFTWARE PRECERTIFICATION PROGRAM: A WORKING MODEL (January 2019), available at https://www.fda.gov/medical-devices/digital-health/digital-health-software-precertification-pre-cert-program.

consider how these benefits compare to the various costs created by software patents. Any fortification of protection for machine learning software should be tied to mechanisms for improving patent quality, including the quality of disclosure contained in the patents.

As for adopters, they should be aware not only of the limits of FDA capacity but also that the term "population health management" is often interpreted broadly, to exempt from regulatory oversight software that may have an impact on individual patients. Thus, independent of the FDA, adopters should require, and scrutinize carefully, summary information on training data and process. They should also generally require evidence of performance on data sets entirely different from the training data. Studies that are independent, rigorous, and representative of demographic and other subgroups are particularly critical if the training data is not itself rigorously collected, labeled, and representative. At a minimum, if adopters have not required evidence of such performance, they should monitor prospective software performance vigorously.

Finally, state law tort regimes should be structured to give developers incentives to promote information flow, even in cases where some secrecy over training data and process must be maintained. For example, the provision of independent, rigorously developed performance information broken out by demographic and other subgroups might bolster a defense against state tort law claims that a ML-CD product was designed defectively. Clear warnings about intended use and limits surrounding the certainty of recommendations could bolter defenses against a failure-to-warn claim.

Notably, these defenses would provide incentives for developer accountability without requiring disclosure of key trade secrets. More generally, a legal environment in which developers who share more information are less subject to tort liability flows logically from one

of the most basic principles of tort law – that responsibility for harm associated with a product should be allocated according to ability to avert the harm *ex ante*.

The sum total of these pragmatic adjustments to existing regimes of information flow should move the needle on accountability without compromising innovation incentives. Our results indicate that the current ecosystem has not failed. But it could be adjusted to do significantly better. Such adjustment may be particularly important given the potential future emergence of highly risky, but also highly beneficial, ML-CD models.

Appendix A

Interview Questions

[IRB-approved introduction outlining goals of project]

For VCs and developers

1) When you choose to invest (VCs) or develop (developers), what do you see as major sources of competitive advantage?

2) How does the investment or development environment in this arena look relative to traditional medical devices? Non-medical software?

3) What are your views on the role of patents in stimulating investment?

4) What are your views on how FDA regulation in this area is operating?

5) What kinds of data or model details do adopters want to see?

6) What do you see as the most significant impediments to market adoption?

[If it doesn't come up in answer to #6, ask about tort liability]

For adopters

1) What experience have you had with ML-CD?

2) What information about ML-CD is most useful to you? (How do you think about information requirements relative to risk?)

3) For your institution, what do you see as major potential advantages and disadvantages of ML-CD?

4) Do you have an opinion regarding how FDA regulation in this space is working?

5) What do you see as the most significant impediments to market adoption?

[If it doesn't come up in answer to #5, ask about tort liability]