Will Data Science in Healthcare Overcome Its Many Roadblocks?

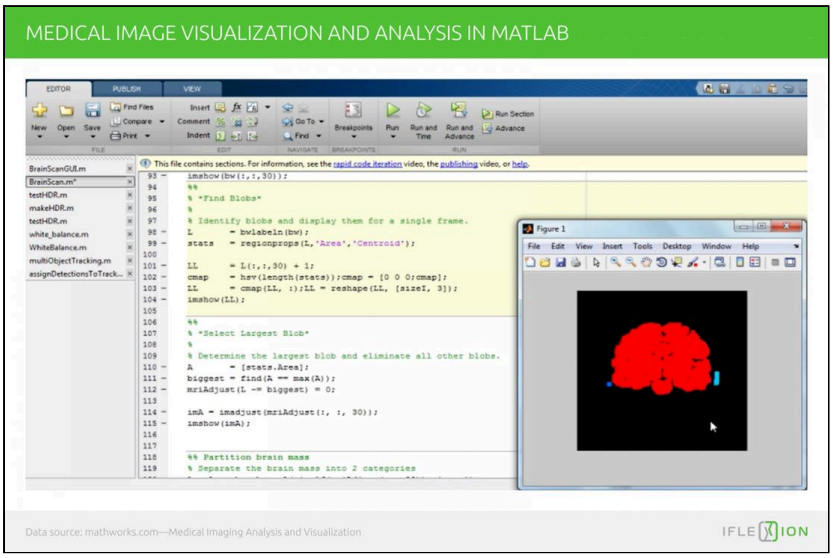
Originally published at https://www.iflexion.com/blog/data-science-healthcare

Published: December 4, 2020 | By Martin Anderson

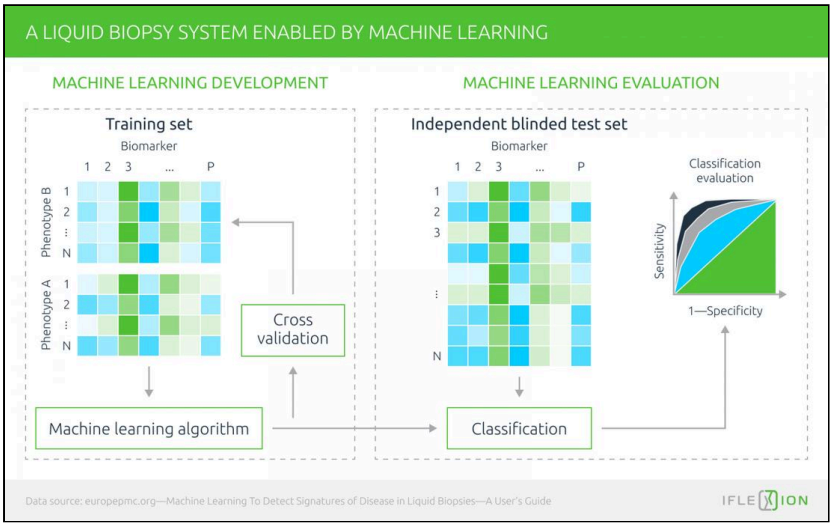
The potential for AI development to transform global healthcare culture from one of 'reactive disaster management' to one of 'prevention and early intervention' through predictive systems has become clear over the last ten years. The potential benefits remain compelling[1,2], with the healthcare AI market forecast recently to reach US$51.3 billion by 2027[3].

Possible applications and developments include:

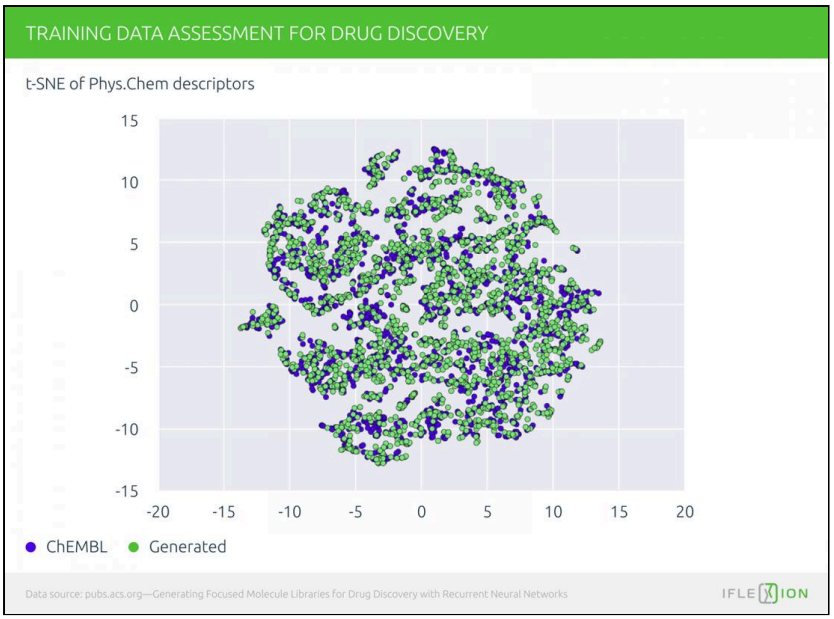
- Automation of medical imaging analysis[4].



- Identification of molecular biomarkers for disease, without biopsies, from simple fluid samples[5].



- Predictive health analytics (PHA) powering AI-based home monitoring systems[6,7] for terminal and high-risk patients (a group representing just 5% of patients but 60% of healthcare costs in the US[8]).
- Drug discovery, including prediction, design and drug-target interaction (DTI) phases[9].



- Machine-driven analysis and logistical strategies for co-morbidities in chronic disease management (CDM)[10].
- Clinical trials for the identification of suitable candidates[11]; for the identification of biologics[12] (large biological molecules) that may contribute to drug evolution for trials; and for ROI analysis[13].

- New machine learning insights and applications related to epidemiology, including statistical prediction models[14] and containment[15].
- The use of AI input for virtual, augmented and mixed reality (VAMR) in healthcare research, monitoring and palliative care[16], as well as for adding an 'informed' VR layer to telesurgery procedures[17].
- AI in genomics research[18,19].
- AI-informed chatbots and engagement systems in the telehealth sector[20].

Systems and applications like this require access to patient data — perhaps the most politically and socially volatile topic in debates over the future of data science and machine learning. Since this represents a major impediment to progress[21], and since it tends to dominate the discourse, let's examine the obstacles, and consider possible remedies.

The Fragmented State of Healthcare Data

The World Health Organization has acknowledged[22] the problems of scale and cost in transitioning healthcare data from paper to digital. Besides highly politicized issues around privacy (which we shall examine), the budgetary demands and long-term commitment needed to fulfil digitization strategies face a number of challenges:

- The limited length of political administrations under pressure to achieve shorter-term victories.
- Sporadic political manipulation of healthcare data[23,24,25], which can render the data less accurate or representative, and adversely affect budget allocation for healthcare IT.
- The limited fiscal resources[26,27] of developing countries.

- The way that US patient data is sub-divided between healthcare insurers[28].
- Endemic undercounting of cases, including undercounting of COVID-19 infections[29,30], which can also skew the data for potential machine learning applications.
- A baffling array, in many countries, of diverse and incompatible digital healthcare solutions across states, counties, and other segmentations of the population[31].
- Resistance from the medical community related to fears about patient outcome under the burden of substandard EHR[32]; fears about the increased security risks of digitizing healthcare infrastructure[33]; a general reluctance, personal and professional, to cede control to machine systems[34].
- Poor provisioning of EHR solutions in competitive, grant-hungry and bid-driven markets[35,36,37].

The United Kingdom's ambitious plans for full patient health data digitization by 2020[38] resulted in only 10% of NHS trusts achieving paper-free records by 2019[39]. This disappointing result came in the wake of the abandoned £12.7-billion National Electronic Health Record Program in 2012[40], possibly the most expensive government IT failure in the world at that time[41].

Around 2009 in the US, a $30-billion government stimulus package for healthcare digitization[42], which might have cohered the country's fragmented health IT sector, was hijacked by the free market: a competitive commercial environment for electronic health records emerged, with rivals eschewing open standards in the hope of market capture through proprietary formats. The result was a multiplicity of non-interoperable EHR systems across US hospitals[43].

Although poorer countries face economic barriers to the costs of implementing digitization, the long-term cost reductions are desirable[44]. In Eastern Europe, where struggling economies make healthcare provision challenging, grants from the European Union are aiding digitization, most notably in the cases of Estonia[45] and Poland[46].

The Outlook for Healthcare Digitization

The political aspect of healthcare data-gathering and disclosure is difficult to solve, particularly where it reveals inequalities. If the route to saving lives and money with machine learning is via naked disclosure of real-world healthcare data/statistics (which may upset governments' more favorable interpretations), the key to progress may be more diplomatic than scientific in nature. Misleading government management of healthcare data systems and statistics predates the pandemic[47,48] and, sadly, seems likely to survive it.

Meanwhile, competitive market forces fragment EHR deployment into multiple non-interoperable networks, particularly in the United States, where federal public health IT proposals must also endure the arguable stigma of 'socialized medicine'[49,50]. In a culture where state/private partnerships have co-opted state provisioning[51], the retarding effects of the free market on healthcare data digitization seem likely to continue after COVID-19, particularly in the US.

If a single large country cannot standardize an EHR system, a global standard seems even more unlikely. The best solution may be cross-systems translation matrices that use robotic process automation (RPA) to extract and homogenize variously-sourced data for the purposes of training AI healthcare and medical research models. Federated solutions may also help, as we'll see later.

If a single large country cannot standardize an EHR system, a global standard seems even more unlikely.

Data Governance, Privacy and Security in Healthcare Data Science

Public ambivalence and consumer fears[52,53] regarding the growing digitization and sharing of health data center mainly on the possibility that health insurance companies will somehow obtain privileged patient information in order to raise costs for coverage against particular conditions or pre-existing genetic markers in individual patients — or even for broader geographical swathes of patients, where research identifies specific health issues in a neighborhood, town or state[54,55,56].

Healthcare's Move from Actuarial to Social Science Data

While insurance rate increases can be explained and justified with actuarial tables, the black-box nature of machine learning presents a novel legal obstacle.

Healthcare companies are rising to this challenge: a new study[57] from the US uses the Python package SHAP to break down the logic of healthcare premium price rises that are calculated with machine learning methods.

Another machine learning project[58] can also demonstrably identify potential high-cost claimants (HiCCs) at a 91% accuracy by scraping publicly available data in an explicable and accountable fashion.

That healthcare insurance providers are keen to switch from actuarial to user-specific data, and that they are willing to use data mining and other oblique techniques to justify increased premiums, has become clear in recent years.

For instance, it was revealed in 2014 that patient information in the UK was repeatedly sold to 178 private health insurance companies[59], leading to a change in the law[60] but also long-term financial implications for those affected.

Two years later it was discovered that the Google/DeepMind AI project had obtained a far wider access to the data of millions of UK patients than had been anticipated in an earlier deal with the UK government[61].

Companies in the US openly vaunt their ability to merge information from data brokers with actuarial data, in order to identify individual health data and case histories[62], while insurance companies are publiclymoving from actuarial to socially mined data via providers such as LexisNexis[63].

The migration from actuarial data is occurring on the pretext of improved early intervention[64], health risk prediction[65], and fraud prevention[66]. However, research has found[67] that these legitimate business uses of web-scraping and other mining techniques can lead to the leveraging of personal health information (PHI) for the purposes of raising consumer premiums.

Gaining Healthcare Data by Stealth and Opportunism

Where anonymization is used in health databases, re-identification can be achieved by cross-referencing the database with another database (such as a voter register) created for a different purpose. This is similar to the way that advertising companies track web-users across domains[68] to facilitate targeted marketing.

Two such databases may each be legal and legitimate in scope, regulation and intent; but cross-referencing them can enable re-identification of anonymized data[69] — a use likely not permitted under the acquisition terms of either dataset.

In the United States and elsewhere, legislation against such abuses has proved easier to implement than enforce. In the US, the HIPAA act of 1996[70] and the HITECH Act of 2009[71] were conceived, respectively, to address governance and security concerns regarding the use of PHI in machine systems and networks.

In Europe this issue has been covered[72,73] by the GDPR since 2018, and with varying levels of governmental commitment in other frameworks around the world[74] over the last few decades.

Healthcare Security Breaches Disarm Legislation

The reach of such laws has been undermined not only by the aforementioned impediments to digitization, but the way that the digitization process massively increases the attack surface for healthcare infrastructure:

- Public healthcare institutions in the US are enduring year-on-year waves of ransomware attacks[75] as the rate of digitization and network-facing data systems grows, with the U.S. Cybersecurity and Infrastructure Security Agency anticipating a greater rise in incidents[76].
- A DNA testing service had 92 million accounts breached in 2018, leading to fears[77] that the stolen data could end up in the hands of insurance companies, if only because publication puts it into the public domain and into the path of insurers' web-scraping and data mining activities.
- The U.S. Department of Health and Human Services Office for Civil Rights currently lists 686 data breaches of health-related companies[78]. The list only reports the previous 24 months, and only covers the United States.
- In 2017, a report on cyber-incursions on healthcare providers detailed the sale of PHI on the deep web, and the common exploitation of unsecured medical IoT devices and networked printers in health-related environments[79].

Stolen health data is frequently posted into the public domain, either as punishment for non-payment of ransoms, or as an anarchic act. Mismanagement and poor security practices can also lead to inadvertent publication of PHI[80,81,82]. In all cases, there's no way to repair the damage done.

Possible Solutions for Healthcare Data Governance, Privacy and Security

Governmental Mandates and Legislation

In the US, there are currently no federally mandated security procedures for healthcare or other sectors[83]. While stronger laws around the world (such as the GDPR) puts companies at greater default liability for poor security practices, the best of them do little more than insist on secure encryption protocols. However, some governments actively seek to undermine even this minimal safeguard[84,85,86].

In fact, social engineering and poorly-designed network infrastructure, among other factors[87], frequently play a more important role in healthcare data incursions[88].

Internationally and on a state-by-state basis in the US, varying statutory provisions exist to limit price optimization by insurance companies, among the most effective being California's Proposition 103 (1988)[89]. Such legislation enshrines the actuarial approach against any better and later technologies (including machine learning), and enforces this for social and economic reasons.

Clearly the quality and scope of legislation must be updated to address new issues arising from machine learning: in Sweden, the Analytic Imaging Diagnostics Arena (AIDA) has recently begun to develop a new legislative framework[90] to address healthcare data sharing in the AI era, while a similar proposal out of Stanford[91] attempts to clarify the extent to which patient data can be commercialized.

New initiatives like this would seem to be a necessary step for all countries that currently rely on older laws to govern newer technologies.

Siloed Institutions for PHI Processing

Air-gapped networks[92] are often used by sensitive industries such as nuclear power, air traffic control, and the military, though it's not an unsolvable problem for determined hackers[93] and does not remove the possibility of social engineering attacks.

However, with appropriate funding and popular support, this approach can work: the private/public partnership behind the Danish National Genome Center (DNGC) sets a high standard for state governance of sensitive medical data: only one copy of a patient's genetic data exists in the center[94], which cannot be copied or erased and will never be allowed to cross borders[95].

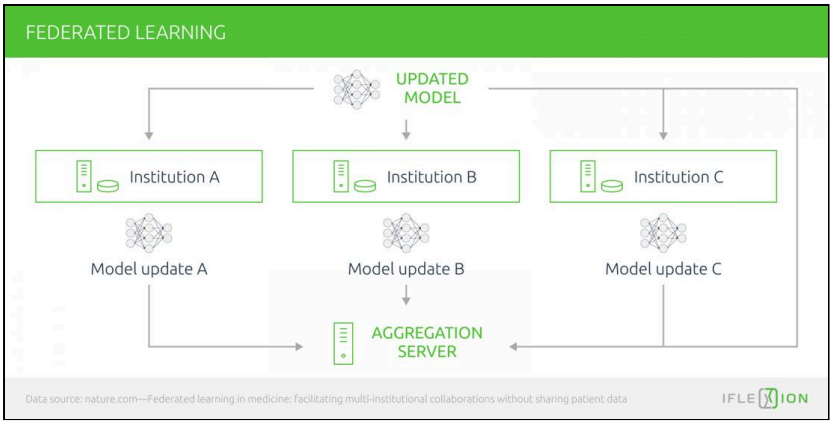
The DNGC is not just a 'data center' for PHI, but operates many research initiatives under the protection of the same level of state-led governance that ordinarily encompasses tax information, driving licenses, and other civic functions precluded from the private sector.

That said, transferring this concept from a small country like Denmark to a larger one like the USA would entail the creation of a well-funded organization on a scale equivalent to the CDC — with a predictable accompanying gauntlet of political and public objection around costs and the de-emphasis of private sector participation.

Machine Learning Approaches

In theory, machine learning can solve its own problems in this respect[96]. In July 2020, new research into federated learning proposed a system where multiple institutions contribute data 'blindly' into machine learning models[97]. Under this approach, all participants benefit from the entirety of

the results without gaining access to the other contributing databases.



In this case, the black-box nature of neural network data models serves as an aide to security, since the resulting algorithms will be too abstracted to deconstruct.

Other AI-based approaches include:

- PrivPy[98], a secure multi-party computation (MPC) framework for machine learning tasks.
- SecureML[99], a system of protocols to preserve privacy in machine learning tasks.
- An edge-based framework to prevent centralization of healthcare data on the cloud[100].

Conclusion

Left entirely to market forces, the evolution of AI-driven machine learning in the healthcare sector risks to reduce global healthcare availability through increased premiums, and turn the known social benefits of affordable coverage into a genetic lottery, among many other potentially negative effects.

However, the technologies that enable this arguable abuse of knowledge also have the potential to prolong and improve life. Society must navigate and mitigate this conflict through legislation that balances a reliance on market forces against the hard-won tenets of civilization. Though many of the specific challenges around this are technical, the central problem is ethical and sociological in nature.

The central problem of the AI evolution in healthcare is ethical and sociological in nature.