Revitalizing pharma's drug discovery pipelines

By **Maria Chatzou Dunford, PhD, CEO and co-founder, Lifebit**

Pharma's Productivity Pressing Needs: Resolving the issue of data accessibility and usability

Life sciences companies have spent many years and resources fine-tuning their drug development processes. Now that the federated data era – a model which combines distributed data sources into a holistic view and applies analytics technology to develop deeper insights – has matured, the next great leap in drug development will come from clinico-genomics data. Is the industry positioned to take full advantage of the latest technological advancements? Innovative companies, such as London-based Lifebit, deliver a more sustainable approach to handling distributed biomedical data sets by leveraging its patented federated technology. Data can be analyzed as if it were in one place without ever having to move it and risking data security. Ultimately, this eliminates fragmented data and unlocks access to the world's distributed biomedical data and its great potential.

As these distributed biomedical data continuously multiply, pharmaceutical companies must find more efficient ways of accessing and working with these large volumes of complex data. Today, research and development (R&D) teams face added work to make sense of siloed data sets, as they cannot easily combine and analyze these distributed data in a seamless and efficient way. Organizations that uncover how to leverage the volumes of clinico-genomic data they currently possess could dramatically boost R&D productivity and gain a major competitive advantage.

In response to the challenge of connecting and accessing distributed datasets, Lifebit leverages its patented federated technology to create a virtualization layer that elastically shrinks and expands, on demand, as access requests to the data are made. In turn, this gives Lifebit's solution the flexibility to connect distributed datasets without having to move the data into one centralized location, which puts sensitive data at risk of interception by third parties.

Partnerships are key to create and share large data sets

Population genomics initiatives led by genomic centers and driven by technology developments have flourished over the last decade. These centers install, maintain and update sequencing

instruments, sample preparation platforms, and computer systems that meet the significantly increased demand for data year after year. For example, international academic consortia in oncology, including the Cancer Genome Atlas (TCGA), International Cancer Genome Consortium (ICGC), and Pan-Cancer Analysis of Whole Genomes (PCAWG), have leveraged such technologies to sequence large numbers of patients to improve the detection of relevant cancer targets.

In line with these initiatives and the growing number of successful genomics-driven drug launches, innovative pharmaceutical companies are now focusing investments into genomics to refuel drug discovery. AstraZeneca, for instance, established the 2 Million Genomes Project, which is a massive effort to compile genome sequences and health records over the next decade.[5] GlaxoSmithKline has also recently invested £40M to expand its partnership with the UK Biobank, supporting the sequencing of the 500K individuals and generating a uniquely rich data resource that is anonymized and secure.[6] These large drug companies are hungry for data, and rather than sourcing it all themselves, they are tapping into population data sets that are rich in genomic, phenotypic, and clinical data.

By recognizing the importance of translating genomic discoveries into healthcare and disease prevention, countries have begun to launch national health data collection initiatives that bring together many stakeholders, including government, medical and research communities, and industry. These initiatives are growing – in fact, by 2025, more than 60 million patients are estimated to have their genomes sequenced in a healthcare context, harvesting more than 12 Exabytes of clinico-genomics data for pharmaceutical companies.[7]

Many countries are now part of the '100K (and over) Genome Club,' including the United Kingdom (Genomics England and UK Biobank), the United States (All of Us), Japan (Initiative on Rare and Undiagnosed Diseases), Saudi Arabia (Saudi Human Genome Program), and Estonia (Personalized Medicine Program). Each initiative is established with the goal of collecting genotypic, phenotypic, and clinical data from a variety of volunteer participants. For instance, Genomics England focuses on collecting whole genomes and associated clinical data from patients with cancer and rare diseases, plus their families. In contrast, the other leading UK initiative, UK Biobank, collects data from primarily healthy volunteers. Other initiatives, such as All of Us, are committed to enrolling volunteers of many different races and ethnicities to address the lack of diversity in genomic data which is a longstanding barrier to translating precision medicine research into real-world practice. All these groundbreaking initiatives enable researchers to improve disease diagnosis and identify drug targets to develop precision medicines.

Impact of genomics on drug development: the demonstrated power of data in drug discovery and development

Most drug failures have been attributed to an incomplete understanding of the link between the biological target of a drug and human disease. To the point, it has been argued that drugs developed with human genetic evidence are more than two times as likely to be approved and eventually make it to patients compared to historical approaches.[1] Furthermore, genomics enables a more personalized approach to drug development, leading to safer and more effective drugs, helping to prevent the 197,000+ deaths each year in Europe caused by adverse drug reactions.[2] In fact, genomics has been used increasingly by drug companies, and several recent successes have demonstrated the efficacy of genomic data to predict the success of new drug targets.

"The original Genomics 1.0 approach is hitting a wall. Moving data is risky and the sheer size of the data adds other logistical concerns. Genomics 2.0, with a technology-driven, federated data approach, is a better, safer solution."

Dr. Anne Deslattes Mays,
Head of R&D at Science & Technology Consulting LLC

The power of generating and analyzing genomic data has resulted in accelerating drug discovery – a 10-fold increase in data corresponds directly with a 100x increase in findings.[8] One pharmaceutical company reported a 50% increase in regulatory submission approvals since implementing genomics as part of its FDA submission process. As such, leveraging the enrichment of population-scale biological and clinical data through national genomic initiatives will enable the industry to dramatically accelerate the discovery of therapeutic targets and drugs for specific diseases and clinical scenarios.

A notable example is the discovery of gain of function mutations in PCSK9 that cause familial hypercholesterolemia and an increase in cardiovascular risk, which led to the launch of both evolocumab (Amgen) and alirocumab (Sanofi and Regeneron).[3] Another example is the launch of vemurafenib (Genentech and Plexxikon – now part of Daiichi-Sankyo), the BRAF inhibitor, which was the first FDA-approved treatment to demonstrate objective responses for 48.4% metastatic melanoma patients that had the commonly observed V600E BRAF mutation.[4]

Genomics 1.0 has hit the wall

Genomics 1.0 (small sets of data in a centralized location for analysis) no longer works for the size and complexity of today's data sets – and matters will only get worse as datasets continue growing.[9] In response, pharmaceutical companies are adopting new technology built to handle exponentially more diverse and complex genomics datasets.

Genomics 1.0 approaches have become expensive, redundant, and slow, all of which will become a multiplying problem for multi-billion-dollar companies. Furthermore, Genomics 1.0 vendors require clients to hand over their data and place it into a centralized environment. As a result, Genomics 1.0 approaches cannot accommodate joint analysis over distributed datasets, as the data must be located in the same environment in order to be analyzed together. Essentially, the client's data is no longer in their control which poses significant data security risks.

Besides these inefficiencies, strict national regulatory frameworks differ from country to country, further impeding progress. For example, sensitive biomedical data are often restricted from leaving secure environments as this increases the risk of interference by third parties. In Europe, more than 50% of member states do not allow their citizens' genomic data to cross borders, creating significant logistical hurdles for companies partnering with national initiatives.[10] Consequently, large-scale genomic data migration becomes unfeasible.

Fortunately, pharmaceutical companies are turning to innovative solutions to enable the analysis of distributed data where it sits, avoiding data movement across borders. Data silos become a thing of the past. G*enomics 2.0* (see **Figure 1**), the new era of biomedical data accessibility, is swiftly gaining ground globally. These solutions leverage federated technology so researchers can easily access, explore, collaborate, and analyze distributed datasets without movement.

Genomics 2.0 – A Required Shift for Today's Increasingly Complex Datasets

As companies continue financially supporting population genomics initiatives to add more data to their exponentially growing, distributed collections, it will become increasingly difficult to access these datasets scattered across the globe in an effective and compliant manner. The answer is Genomics 2.0.

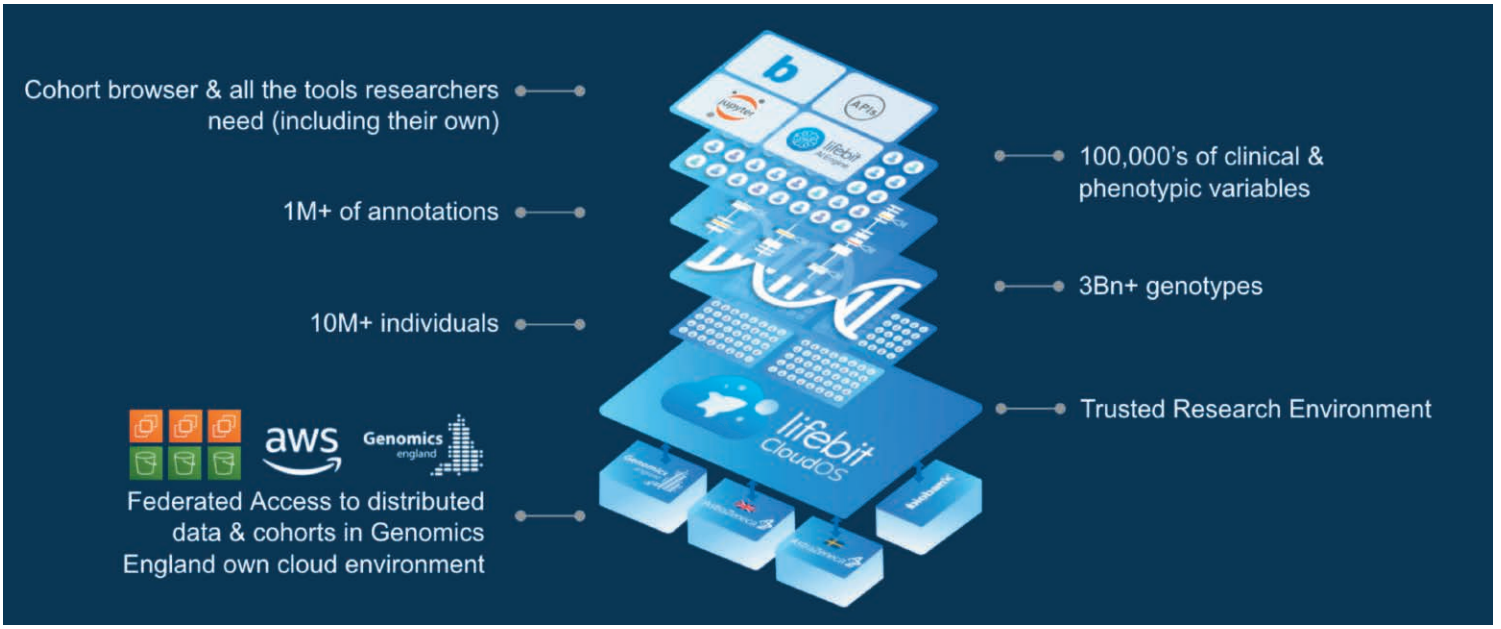

Figure 1: The new era of biomedical data accessibility

Lifebit CloudOS brings a technological guarantee and agility that Genomics 1.0 approaches cannot provide, advancing infrastructure as a key enabler for Genomics England researchers.

Lifebit was founded to disrupt the global pharmaceutical industry and population genomics initiatives with innovative federated and advanced AI solutions with *Genomics 2.0*. One early adopter of *Genomics 2.0* is Genomics England which has recently partnered with Lifebit to launch a federated Trusted Research Environment, offering pharmaceutical companies access to their rich datasets as a read-only library, rather than a lending library.[11] Consequently, data always remains within Genomics England's control, and researchers can freely access, browse, and analyze the data without transfers.

"Researchers want access to critical, but inaccessible information stored in the world's hospitals, biobanks, and R&D centers. Unlocking the ability to access this global data through federation will be the next frontier in accelerating drug discovery and AI powered insights."

Dr. Anne Deslattes Mays,
Head of R&D at Science & Technology Consulting LLC

Case Study: Genomics England Overcomes Data Accessibility and Security Challenges

The Trusted Research Environment is owned by Genomics England (GEL) and managed by Lifebit, thereby avoiding vendor lock-in, as Genomics England is always in possession of their environment and, consequently of their data. Unlike Genomics 1.0, Lifebit's platform solution, Lifebit CloudOS, is never deployed within Genomics England's cloud environment and does not require any handover of data. Rather, Lifebit CloudOS acts as a separate layer on top of Genomics England's cloud environment, enabling it to orchestrate and manage the environment without directly accessing data for an additional layer of security.

Federation allows modern solutions like Lifebit CloudOS to create temporary virtual links between Genomics England's dataset and other external datasets to virtually combine the distributed data and run joint analysis. The same federation principles can also link code repositories to Genomics England's Environment to introduce new tools and workflows without exposing IP, thus allowing researchers to use their preferred tools instead of relying on a pre-populated list of standard tools and workflows. Once complete, the virtual links are terminated, and no further communication is authorized. Data and code never move, preventing cross-border data movement and intellectual property (IP) issues.

One of the features of the Lifebit platform is that researchers can analyse their in-house alongside Genomic England data, without ever moving the data into the Genomics England environment. Researchers' in-house data do not intermingle with GEL patient data, which protects Genomic England data sets from potential malware or spyware, and guarantees data security and integrity. By using federated links, the Lifebit solution enables researchers to leave their data in-situ and operate as if the data were in one place – which is both more secure, and more efficient. This achievement is important in itself, but also points to a future where we could federate genomic datasets internationally, a key theme that Genomics England are looking to drive with Lifebit via The Global Alliance for Genomics and Health (GA4GH).

To facilitate the identification of cohorts for researchers to run their analyses over, it is critical to have a solution that seamlessly queries the scale and complexities of Genomics England data, including large data sets of genomic, multi-omic, medical records, and other structured data. Genomics England's federated cohort browser can scale to 10M+ individuals and more than three billion genotypes by leveraging Lifebit's technology which minimizes the amount of necessary storage for data, and the time required to run different queries over the database due to the extremely efficient indexing and bit operations capabilities. Importantly, the cohort browser will be able to scale as Genomics England's collection grows over time to include more data points and types (see **Figure 2**).

Essentially, by choosing a *Genomics 2.0* approach, Genomics England has solved both the economic inefficiencies involved with moving big data while remaining compliant with data policies regarding sensitive patient data.

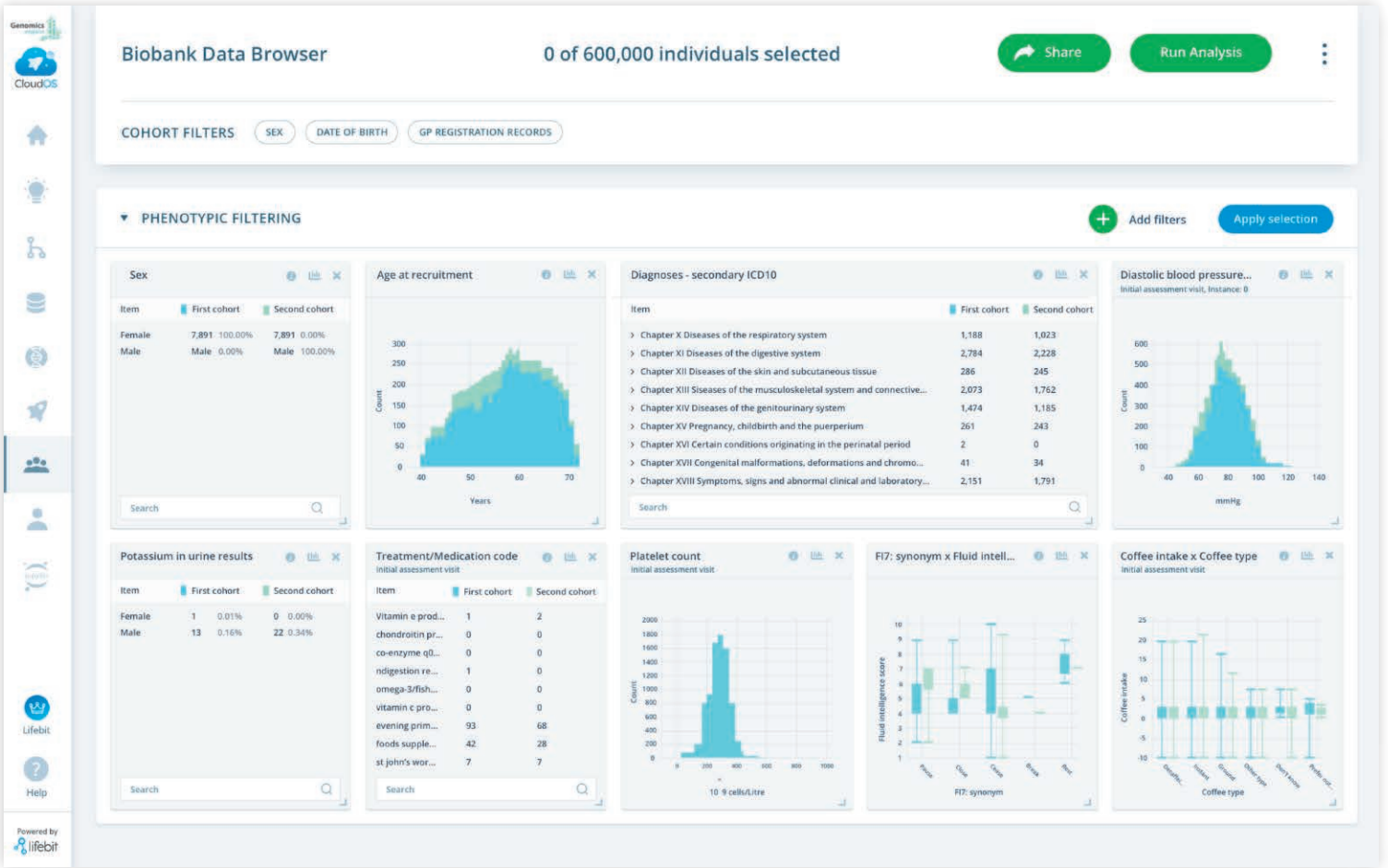

Figure 2: Lifebit CloudOS' Scalable and Federated Cohort Browser

Built to handle billions of genotypes, hundreds of thousands of clinical phenotypic variables and millions of annotations. This provides a seamless user experience enabling R&D teams create, share and analyze cohorts of interest in seconds.

Genomics England's approach to distributed analysis offers a blueprint for others to unlock access to the world's distributed genomic data, as this solution is easily reproducible in other environments.

Dissolving Pharmaceutical Data Silos with Data Exchanges

Estimates reveal that 80% of companies do not have an integrated strategy for data management, leading to distributed and inaccessible data.[12] As noted above, data silos are an impediment to the efficiency and performance of research and development teams. Although pharmaceutical companies are heavily investing in obtaining clinico-genomics data, it remains difficult and time consuming to identify or generate the right data (*de novo* or complementary). As a result, there is a significant gap (and opportunity) in bridging the potential value of the data sitting in distributed data lakes and the ability for pharmaceutical R&D teams to optimize its value in drug development and other R&D activities.

"A key theme that Genomics England are looking to drive with Lifebit via The Global Alliance for Genomics and Health (GA4GH). "This is clearly the direction of decentralised genomic research for the future."

Parker Moss, CCO, Genomics England

To close the gap, companies have been leveraging data exchange solutions to make data visible, accessible, and usable for researchers to increase its value. These innovative data exchange platforms require sophisticated federated technology to unlock the door to distributed clinico-genomic data, allowing data custodians to showcase valuable data summary statistics. Data customers, on the other hand, rely on federated technology to browse summary statistics, identify, and build cohorts of interest for their research.

By changing the current data landscape, scalable data ecosystems (such as Lifebit's) promote collaboration and distributed data ownership. A centralized network of data identification and access ensures that researchers have all the data they need to run complex analyses by enabling access to the world's clinico-genomic data in a streamlined and structured way.

Integrating Multiple Data Sources to Derive Deeper Insights

In addition to ironing out data accessibility issues, pharmaceutical companies need to transition from big data to deep data by placing raw data into context to derive deeper insights. Genomic data is no longer sufficient on its own for companies to remain competitive. The value lies in the integration of all types of data, including clinical, demographic, genotypic, phenotypic, and real-world data.

In recent years, drug development has made its way from the traditional pharmaceutical labs to the real world, where companies can access data

collected during routine patient care (i.e., Electronic Health Record (EHR) data), which is generally more comprehensive than data collected during tightly controlled clinical trials. Furthermore, real-world data is representative of a much larger cohort of patients that companies can potentially have access to – it's estimated that less than 5% of all cancer patients enroll in clinical trials.[13]

The challenge with real-world data, however, is that it's often locked in the location where it is generated: a patient's EHR data generally stays within their healthcare provider's software system, whereas their sequencing data remains within the sequencing laboratory's compute infrastructure. Connecting these distributed data sets is essential for drug companies to improve patient stratification for more successful clinical trials that are representative of a diverse population, especially as precision medicines are developed based on specific biomarkers.

To create comprehensive longitudinal databases, pharmaceutical companies have started to turn to federated approaches as no other approach enables them to link distributed data in an efficient way while leaving data at its source. Lifebit's solution creates temporary, virtual links, as it does for Genomics England, to integrate different data types. As such, consistent integrated longitudinal databases can be virtually created to enable R&D researchers to submit complex queries to the federated system, which effectively combines data from multiple sources of different types, even if each individual source does not possess all the functionality needed to answer the query. In addition to querying complex data, researchers can seamlessly run joint analyses over distributed data sets, whether in-house or external, increasing the power of their analyses.

Artificial Intelligence (AI) can further increase the depth of data collections when combined with federated technology. AI algorithm performance significantly improves when large troves of data can be accessed through federated links. By enhancing data integration approaches through federation and AI, R&D teams can delve deeper and generate more meaningful insights, optimize patient recruitment for clinical trials, develop a better understanding of various therapeutics, and directly improve patient outcomes.

Future vision for accelerated therapeutic breakthroughs

Pharmaceutical companies are deploying a host of strategies to face the increasing competitive pressure from smaller biotechnology and larger technology companies. To remain competitive, some have started to leverage the exchange of intellectual property and R&D expertise by creating alliances with biotechnology companies and academia: for some time now, pipelines within the top 30 pharmaceutical companies are sourced through external innovation.[14]

This is best exemplified by the current pandemic and the race to develop a vaccine for the SARS-CoV-2 virus. The three leading and approved vaccines (at the time of this writing) have been developed through industry and biotech/academic collaborations and include the Pfizer-BioNTech (BNT162b2), Moderna-NIH (mRNA-1273), and AstraZeneca-Oxford University (AZD1222) vaccines.

Genomics took center stage and facilitated the rapid development of these vaccines, as two vaccines were generated using novel messenger RNA (mRNA) technology which instruct cells how to make coronavirus spike proteins (Pfizer-BioNTech and Moderna-NIH).[15] Furthermore, collaborative efforts and enhanced access to global COVID-19 sequencing data enabled pharmaceutical companies and their collaborators to develop vaccines in a fraction of the typical development timeframe. Industry should factor these recent learnings into future R&D processes. It's clear: genomics and clinical data are enablers of innovation and a potential solution to common innovation bottlenecks currently hindering drug development speed. Due to the agile R&D processes industry adopted, many governments are now vaccinating their population against COVID-19 less than a year after the first case was reported in Wuhan, China.

Pharmaceutical companies are rethinking their drug development processes. Federated and AI-driven platforms, such as Lifebit's, offer a more sustainable and innovative approach to big data with its federated AI solution, where data is analyzed globally and comprehensively by eliminating fragmented access to data, and unlocking access to the world's distributed biomedical data. JPM

Dr. Maria Chatzou Dunford

Maria is the CEO and co-founder of Lifebit. Maria, is a thought-leader and biotech innovator, expert in AI-driven drug discovery, biomedical informatics and federated computing. She is also a passionate entrepreneur and has founded two companies, Innovation Forum Barcelona and Techstars-backed Lifebit. Prior to Lifebit, she was a biomedical researcher, working on developing tools and methods that facilitate the analysis of Big Biomedical Data and promote personalised medicine discoveries. This includes the industry's standard programming framework, Nextflow, that has revolutionised the computational analysis of genomic data.

Maria is also a frequent industry speaker and has spoken in many international conferences on the subjects of genomics workflows, the computational challenges of personalised medicine, AI, Cloud and HPC in genomics and drug discovery, women in leadership, entrepreneurship, science ventures, among many other topics.

You can contact Dr. Maria Chatzou Dunford at maria@lifebit.ai

References

1. King, Emily A., J. Wade Davis, and Jacob F. Degner. "Are drug targets with genetic support twice as likely to be approved? Revised estimates of the impact of genetic support for drug mechanisms on the probability of drug approval." PLoS genetics 15.12 (2019): e1008489. See full resource https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1008489.
2. European Commission; Memo, "Strengthening Pharmacovigilance to reduce adverse effects of medicines." (December 10, 2008) See full resource https://ec.europa.eu/commission/presscorner/detail/en/MEMO_08_782.
3. Tavori, Hagai, Ilaria Giunzioni, and Sergio Fazio. "PCSK9 inhibition to reduce cardiovascular disease risk: recent findings from the biology of PCSK9." Current opinion in endocrinology, diabetes, and obesity 22.2 (2015): 126. See full resource https://journals.lww.com/co-endocrinology/Fulltext/2015/04000/PCSK9_inhibition_to_reduce_cardiovascular_disease.10.aspx.
4. Kim, Geoffrey, et al. "FDA approval summary: vemurafenib for treatment of unresectable or metastatic melanoma with the BRAFV600E mutation." Clinical Cancer Research 20.19 (2014): 4994-5000. See full resource https://pubmed.ncbi.nlm.nih.gov/25096067/.
5. Leidford, Heidi, "AstraZeneca launches project to sequence 2 million genomes," Nature News. (April 22, 2016) See full resource https://www.nature.com/news/astrazeneca-launches-project-to-sequence-2-million-genomes-1.19797.
6. GlaxoSmithKline; Press Release, "GSK welcomes launch of the UK Government's Life Sciences Sector Deal." (December 6, 2017). See full resource https://www.gsk.com/en-gb/media/press-releases/gsk-welcomes-launch-of-the-uk-government-s-life-sciences-sector-deal/.
7. Birney, Ewan, Jessica Vamathevan, and Peter Goodhand. "Genomics in healthcare: GA4GH looks to 2022." BioRxiv (2017): 203554. See full resource https://www.biorxiv.org/content/10.1101/203554v1.
8. Visscher, Peter M., et al. "Ten years of GWAS discovery: biology, function, and translation." The American Journal of Human Genetics 101.1 (2017): 5-22. See full resource https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5501872/.
9. Stephens, Zachary D., et al. "Big data: astronomical or genomical?." PLoS biology 13.7 (2015): e1002195. See full resource https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1002195.
10. European Commision; News, "EU countries will cooperate in linking genomic databases across borders." (April 10, 2018) See full resource https://ec.europa.eu/digital-single-market/en/news/eu-countries-will-cooperate-linking-genomic-databases-across-borders.
11. Genomics England; Press Release, "Genomics England launches next-generation research platform central to UK COVID-19 response." (June 29, 2020). See full resource https://www.genomicsengland.co.uk/research-environment-covid-19-lifebit-aws/.
12. Stanford Medicine; White Paper, "Harnessing the Power of Data in Health." (June 2017). See full resource https://med.stanford.edu/content/dam/sm/sm-news/documents/StanfordMedicineHealthTrendsWhitePaper2017.pdf.
13. Unger, Joseph M., et al. "The role of clinical trial participation in cancer research: barriers, evidence, and strategies." American Society of Clinical Oncology Educational Book 36 (2016): 185-198. See full resource https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5495113/pdf/nihms870923.pdf.
14. Deloitte Centre for Health Solutions; "Ten years on: Measuring the return from pharmaceutical innovation 2019." See full reference https://www2.deloitte.com/content/dam/Deloitte/uk/Documents/life-sciences-health-care/deloitte-uk-ten-years-on-measuring-return-on-pharma-innovation-report-2019.pdf.
15. Matloff, Ellen. "The COVID-19 Vaccines that Genomics Built," Forbes.com and reprinted in My Gene Counsel. (November 21, 2020) See full resource https://www.mygenecounsel.com/the-covid-19-vaccines-that-genomics-built/.