Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220601

Augmenting THerapeutic Effectiveness Through Novel Analytics (ATHENA) - A Public and Private Partnership Project Funded by the Flemish Government (VLAIO)

Ploingarm PETSOPHONSAKUL[a,1], Ashkan PIRMANI[b], Edward DE BROUWER[b], Murat AKAND[c], Wouter BOTERMANS[a], Frank VAN DER AA[c], Joris Robert VERMEESCH[d], Fritz OFFNER[e], Roel WUYTS[f], Yves MOREAU[b], Ingrid MAES[g], Ines BLOCKX[a], Patricia VAN ROMPUY[h], Martine LEWI[a] and Bart VANNIEUWENHUYSE[h]

[a] *Janssen R&D/Clinical Innovation, Belgium*
[b] *ESAT-STADIUS, KU Leuven, Belgium*
c *Dept. of Urology KU, Belgium*
[d] *Center for Human Genetics, UZ Leuven, Belgium*
[e] *Leuven, Dept. of hematology University of Ghent, Belgium*
f IMEC, Belgium
[g] *Inovigate, Belgium*
[h] *Janssen BeNeLux, Belgium*

Abstract. The complexity and heterogeneity of cancers leads to variable responses of patients to treatments and interventions. Developing models that accurately predict patient's care pathways using prognostic and predictive biomarkers is increasingly important in both clinical practice and scientific research. The main objective of the ATHENA project is to: (1) accelerate data driven precision medicine for two use cases – bladder cancer and multiple myeloma, (2) apply distributed and privacy-preserving analytical methods/ algorithms to stratify patients (decision support), (3) help healthcare professionals deliver earlier and better targeted treatments, and (4) explore care pathway automations and improve outcomes for each patient. Challenges associated with data sharing and integration will be addressed and an appropriate federated data ecosystem will be created, enabling an interoperable foundation for data exchange, analysis and interpretation. By combining multidisciplinary expertise and tackling knowledge gaps in ATHENA, we propose a novel federated privacy preserving platform for oncology research.

Keywords. Precision medicine, Federated platform, Machine learning, Distributed analytics, Data science

[1] Corresponding Author, Ploingarm Petsophonsakul; E-mail: ppetsoph@its.jnj.com.

1. Introduction

Pharmaceutical compounds and/or other therapeutic interventions may not show the same level of efficacy in all patients, leading to the term "imprecision medicine". Any approach that can enhance the effectiveness of therapy should be welcomed. There is a strong belief in the benefits of enriching available clinical patient information (non-omics data) with omics data which refers to large biomarker data sets characterizing biological features such as genomics, transcriptomics, proteomics etc. [1]. Omics data is considered as supporting data to gain a deeper understanding of the complex multifactorial causes and the natural evolution of a disease; as well as defining patient characteristics that can predict treatment success [2]. Such knowledge is needed for a wide range of diseases, most notably in oncology. Thus, improved disease insight in combination with predictive analytics form the basis of Personalized Medicine.

Project ATHENA (Augmenting THerapeutic Effectiveness through Novel Analytics), aims at gaining a deeper understanding on the various challenges of creating this combined omics and non-omics data approach for researching two cancer types: one solid tumor (bladder cancer) and one hematological cancer (multiple myeloma). Bladder cancer is selected because of its challenges in risk stratification, and multiple myeloma because it is regarded as one of the most complex cancer types exacerbated by a multitude of treatment options [3]. Improvements in targeted treatment will require the identification of relevant, actionable biomarkers to either risk stratify or adapt treatments. Identification of such markers requires analyses of data across the entire pheno-genotypic (omics and non-omics) set of characteristics and across a fully longitudinal follow-up of an entire patient cohort. This can create considerable privacy challenges. To overcome these, we will take a federated and privacy preserving analytics approach whereby data stays local under the control of the original data custodians.

Current state-of-the-art in Omics/Non-omics integration and Federated learning

Omics data integration has been addressed in recent years [4, 5]. However, only a few of them resulted in omics-based algorithms with sufficient predictive ability to be implemented into clinics or public health domains [2,6]. The relatively poor predictive ability of genomic data may be explained by the difficulty to analyse and extract relevant information from the omics data and by the large variation of health-related traits explained by non-omics data [7]. Therefore, it is crucial to integrate omics and non-omics data in the same models [8]. There is strong demand for federated systems that enable joint analysis efforts across multiple partners holding sensitive or competitively valuable data. Several projects aimed to address data privacy including Machine Learning Ledger Orchestration for Drug Discovery (MELLODDY), where machine learning is used to accelerate drug discovery while ensuring privacy preservation of both the data and the models through federated learning [9].

2. Concrete Objectives and Criteria

The main objective of the present project is to generate new knowledge to create a federated privacy preserving machine learning platform that can execute the newly researched machine learning algorithms according to medical case requirements.

2.1. The supporting info/infrastructure will require research on and integration of the following data pipelines:

- Non-omics data processing pipeline for clinical and patient recorded data.
- Non-omics imaging data processing pipeline with algorithms for the standardization and privacy preservation of imaging data.
- Standardized high performance pipelines for omics data processing.
- High performance somatic variant calling pipelines.
- Systems genetics pipeline for feature extraction from omics data.

A modular technical design will be used in support of data gathering and distributed analytics. Data pipelines (clinical, image and omics) and the output of each pipeline comes together in an integrated data warehouse for each of the participating institutions. These data warehouses form the basis for the exploration via distributed machine learning.

2.2. The performance of the platform should enable optimum data flow and linkage with proper user management incorporating appropriate user authentication and authorisation and thus be robust once scaled up.

2.3. Explore, research, and develop

- Novel capabilities for data from real-world patient trajectories.
- Federated and privacy-preserving implementations of these techniques that let consortium partners protect the privacy of patient data and maintain full control over the processing of these data at all times.
- Appropriate algorithms to ensure standardization and privacy protection of imaging data for the creation of the non-omics visualization pipeline.

2.4. Accelerate data driven precision medicine for two use cases – bladder cancer and multiple myeloma.

Bladder cancer: To develop a retro- and prospective longitudinal dataset of the entire population treated for non-muscle-invasive bladder cancer (NMIBC) including phenotype and genotype information, stored on local data sources and capable of generic data extraction to address NMIBC key scientific questions. **Multiple myeloma:** To develop a prospective longitudinal dataset of the entire population treated for multiple myeloma including phenotype and genotype information, stored on local data sources and capable of generic data extraction to address key scientific questions.

2.5. Define a governance framework (incl. legal, ethical, data privacy aspects) and conduct project management to successfully accomplish the project goals.

3. Approach

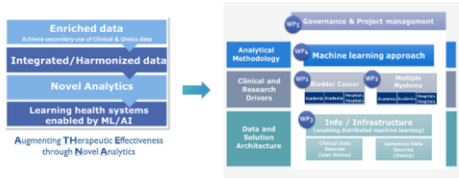

Figure 1. Structural approaches of ATHENA

The research will tackle the fundamental challenges associated with integrating omics and non-omics data by means of investigating and creating the appropriate data ecosystem that will enable an interoperable foundation for data exchange and capability. This will serve as the underlying info/infrastructure to facilitate the investigation and creation of a federated and privacy-preserving machine learning platform. A schematic overview of the ATHENA research project with different work packages (WPs) is provided in Figure. 1.

4. Deliverables (year 1)

1. Standardization of data catalogue and harmonization of NMIBC and multiple myeloma data
2. Independent Privacy/Security audit of local instance of Feder8 platform, a federated data network solution, and implementation of the central and local Feder8 component
3. High-level architecture blueprint for data pipelines integration
4. Approvals of technical and retrospective bladder cancer protocol

5. Impact and Expected Outcome

Resulting machine learning methodologies and the clinical outcomes of the project represent new knowledge that is set to advance the current state of the art in their respective fields. **Industry impact:** New knowledge on infrastructure, pipelines, warehouse data management, integration, security, standardization and privacy algorithms (analytics) etc., offering valuable economic impact for industry partners and opportunity for future development. **Independent research impact:** Proof of principle on new application-driven machine learning methodologies. Clinical outcomes offering new insights into bladder cancer and multiple myeloma for improved patient treatment. ATHENA is the new omics and non-omics approach to personalized medicine. **Clinical impact:** Insights in disease mechanisms and impact of different progression, optimal care and treatment pathways and patient risk-stratification for bladder cancer and multiple myeloma will be obtained by applying distributed analytics.

Within the ATHENA framework, a federated privacy preserving machine learning platform that can execute machine learning algorithms according to specific medical case requirements will be developed. This platform allows multiple hospitals to collaborate and build a common machine learning model without directly sharing sensitive data

(Figure. 2) Each institution remains in full control of its own data and resources. The platform offers privacy by design and complies with General Data Protection Regulation (GDPR) [10]. Expected project outcomes include (1) Removed barriers in ethics, consent and data governance, data quality and interoperability and affordability (2) Strengthened research partnership between academia and industry (3) Established regulatory frameworks and policy, as a basis for value-based healthcare.

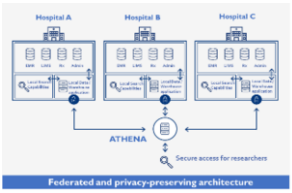

Figure 2. ATHENA privacy-preserving machine learning platform

6. Conclusions

ATHENA brings together a unique, multidisciplinary and complementary partnership of oncologists, experts in IT architecture, data science, high performance cloud computing, genomics and medical affairs. ATHENA utilizes leading-edge technology to integrate and enrich patient level data and analytics. The project leverages machine learning to generate insights from electronic health record (EHR), genomics and medical imaging data, thus creating a federated privacy-preserving machine learning platform that will accelerate data driven precision medicine and provide solutions to ethics, consent and data governance, data quality, interoperability and sustainability.

Acknowledgment

ATHENA is a public private partnership project funded by Flanders Innovation & Entrepreneurship (VLAIO). Project number: HBC.2019.2528

References

[1] Letai A. Functional precision cancer medicine-moving beyond pure genomics. Nat Med. 2017 Sep 8;23(9):1028-1035. doi: 10.1038/nm.4389. PMID: 28886003.
[2] Brandão M, Pondé N, Piccart-Gebhart M. Mammaprint™: a comprehensive review. Future Oncol. 2019 Jan;15(2):207-224. doi: 10.2217/fon-2018-0221. Epub 2018 Aug 29. PMID: 30156427.
[3] Lohr JG, et al. Widespread genetic heterogeneity in multiple myeloma: implications for targeted therapy. Cancer Cell. 2014 Jan 13;25(1):91-101.
[4] Hasin Y, Seldin M, Lusis A. Multi-omics approaches to disease. Genome Biol. 2017 May 5;18(1):83.
[5] Ritchie MD, Holzinger ER, Li R, Pendergrass SA, Kim D. Methods of integrating data to uncover genotype-phenotype interactions. Nat Rev Genet. 2015 Feb;16(2):85-97.
[6] Wallden F, et al. Development of the molecular diagnostic (MDx) DLBCL Lymphoma Subtyping Test (LST) on the nCounter Analysis System. J. Clin. Oncol. 2015;33.
[7] Lichtenstein P, et al. Environmental and heritable factors in the causation of cancer--analyses of cohorts of twins from Sweden, Denmark, and Finland. N Engl J Med. 2000 Jul 13;343(2):78-85.
[8] Gerstung M, et al. Precision oncology for acute myeloid leukemia using a knowledge bank approach. Nat Genet. 2017 Mar;49(3):332-340.
[9] https://www.melloddy.eu/ visited 16/03/2022.
[10] https://ec.europa.eu/info/law/law-topic/data-protection/data-protection-eu_en visited 16/03/2022.