Paper ML05

Artificial Intelligence (AI) Enhanced Precision Medicine Drug Development

Mohammad Afshar, Ariana Pharmaceuticals, Paris, France
Coralie Williams, Ariana Pharmaceuticals, Paris, France
Federico Goodsaid, Ariana Data Intelligence, Cambridge, USA
Jean Sallantin, CNRS LIRMM, Montpellier, France

ABSTRACT

Precision medicine requires new approaches to clinical trial design and analytics. Traditional methods, often requiring large datasets fall short due to data heterogeneity. Therefore, analysis of small datasets with broad heterogeneous data results in significant hurdles. These can be alleviated using advanced Artificial Intelligence (AI) tools to analyze complex datasets in order to identify biomarkers useful for selecting patients for a clinical trial, define surrogate efficacy endpoints or discover new indications for a therapeutic.

Modeling, simulation and innovative statistical analysis can be combined with novel approaches using Formal Concept Analysis (FCA) in Machine Learning. This family of AI tools is starting to be used in modern drug development. However, their deployment requires acceptance across multiple disciplines (e.g. medical and regulatory affairs, bio statistics, translational medicine, etc.) as results impact multiple aspects of drug development.

We will highlight the FCA AI framework with examples from oncology, Alzheimer's disease and diabetes.

INTRODUCTION

Scientific discoveries require the ability to extract hypothesis from very small number of initial observations, and build-up a network of relations that constitute a model of the observed phenomena. It is therefore essential to be able to extract from the data important relations that are non-contradicted, even if their support (i.e. the number of times they have been observed) is small.

The specifics of the analysis "small data" has received special attention [1,2] . For example, the oncology patient repository held by the company Flatiron that covers about 15% of total American cancer patients, converge towards the requirement to analyze small subset of patients: "… we never thought we'd be working with small data sets. And we have found at Flatiron that many of the most important questions we're trying to answer are being answered with 30–60 patients, because these are the subpopulations that have been excluded from clinical trials"[3]. Identification of relations at an individual patient level in small data are essential to "precision health" or "precision medicine.

However, analyzing datasets where the number of observations is orders of magnitude smaller than the number of attributes, which is often the case when handling genomic data, is an analytical challenge for traditional statistical methods. Formal Concept Analysis (FCA) constitutes a Machine Learning technique that is less affected by population size. They can be effectively used to extract and rank in a hypothesis free, data-driven mode phenotypic and genotypic biomarkers from pre-clinical and clinical datasets. We have developed and implemented an FCA framework in the KEM® (Knowledge Extraction and Management) software.

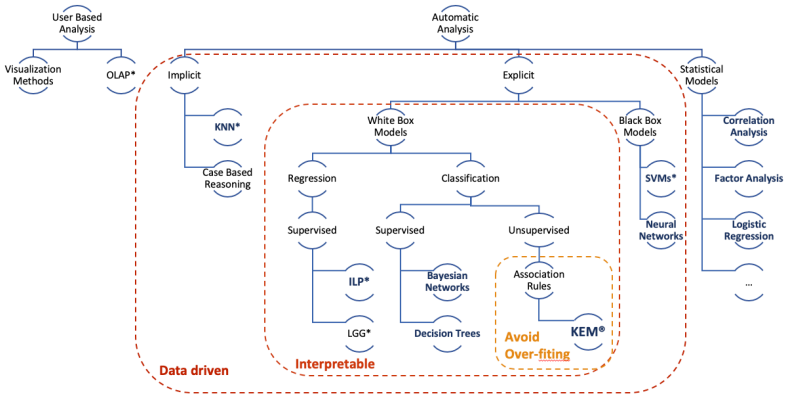
In this paper we will position the FCA AI framework within the space of data analytics methods and illustrate our approach with examples from oncology, Alzheimer's disease and diabetes.

WHICH ARTIFICIAL INTELLIGENCE? GOING BEYOND THE HYPE

A recent analysis by Tran et al[4] explored the impact of AI in health and medicine, as reflected by the number of relevant publications in the field, identified a total of 27,451 papers that were published between 1977 and 2018. This paper illustrates the phenomenal acceleration showing that over 60% of the publications occurred in the last 4 years and 84.6% were dated 2008–2018.

In a very schematic approach it is possible to group statistical hypothesis driven methods away from the data driven approaches that may then be divided into Implicit and explicit models (Figure 1). Within the Explicit models, one can consider analytics that methods such as Support Vector Machine and Neural Networks are harder to interpret ("Black Box") versus other methods that could be represented as "White Box Methods". We note that a large number of biomedical AI publications focus on deep learning and associated "black box" techniques. Interpretable methods may be again sub divided into supervised regression and classification which can be either supervised or unsupervised. Formal Concept Analysis methods mainly belong to the unsupervised group. The Knowledge Extraction and Management (KEM®) software belongs to a sub set of FCA that use a Galois lattice to represent relations.

Figure 1. Schematic representation of Data Analytics methods



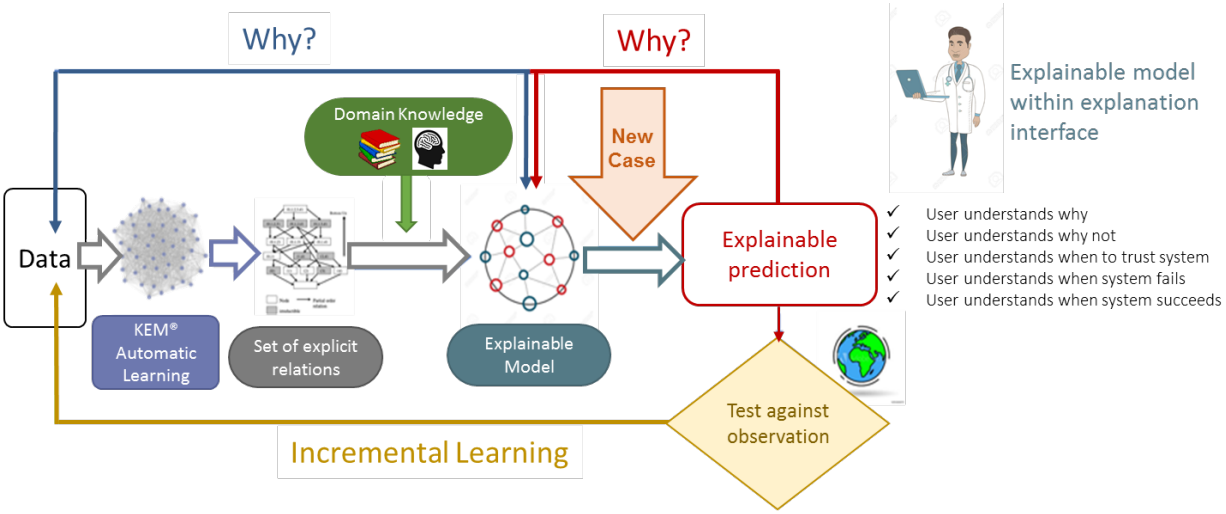
EXPLAINABLE ARTIFICIAL INTELLIGENCE (XAI): FORMAL CONCEPT ANALYSIS AS A FRAMEWORK

Explainable AI (XAI) is likely to become the next frontier of human-machine collaboration[5,6]. A large number of existing AI applications are effectively black boxes, lacking the ability to explain the reasoning behind a decision. Several reports have attributed the limited effectiveness of AI in a number of applications such as use on the military field to the inability to explain decisions and actions to human users[7].

In the healthcare space, broad adoption of AI will require the technology to allow for additional human scrutiny and hence be enabled to provide the line of reasoning leading to a decision. Furthermore, in a number of applications, and in particular in the analysis of data from clinical trials, the information contained in the data represents only a very small fraction of the "field knowledge" that is shared by the scientific and medical community.

It is therefore important for an explainable AI to provide an explicit representation of the knowledge it has extracted from the data which could then be combined by the human experts with additional knowledge from the field into a model.

Figure 2. Schematic representation of Data Analytics methods



This model could then in turn be used for making predictions. Explainable AI won't replace the experts; rather it will complement and support humans, so they can build a more complete model of the system that is analyzed and subsequently make better, faster, more accurate and consistent decisions.

Specifically, incorporating domain knowledge (versus general knowledge) into machine learning and other data-driven approaches has enabled the discovery of novel[8,9] that can be confirmed by more traditional analytical methods.

Formal Concept Analysis provides an ideal framework for XAI. Given a dataset of any dimension and using FCA, the system is able to extract relationships of the general form:

$$X \rightarrow Y,$$

where X and Y can be a unique descriptor or combination of descriptors.

An association rule is a relationship such as, with X being the antecedent (or a combination of antecedents) and Y the consequent (or combination of consequents) of the rule, that allows inferring X as a consequence of Y. Association rules have been developed to detect patterns and relationships in large heterogeneous databases. This methodology was introduced by Agrawal and Sirikant[10] and it has been successfully used in different domains including: market basket analysis, finance, drug discovery, patient stratification and genomic characterization of complex diseases. Association rules may reflect the presence of a causal relationship between two variables. However, the presence of a statistically significant association is not necessary nor sufficient to indicate the presence of such causal relationships.

Extracted *knowledge* is represented by all association rules present in a dataset. This set of rules can be applied to a set of descriptors and propagated to decide.

QUALITY MEASURES

For a dataset of *N* subjects, n_X , $n_y$and n_{XY}, respectively are the numbers of subjects satisfying:

- the antecedent(s) X of a rule
- the consequent(s) Y, and
- both parts of the rule.

For any given rule, it is possible to identify specifically the n_{XY} subjects matching this rule.

Five main quality measures describe and rank association rules:

- Support
- Confidence
- Lift
- Two P-values
- Measures of Significance

Support is the number of subjects for which the association has been observed:

$$Support(X \rightarrow Y) = n_{XY}$$

Confidence is the percentage of the characterized subjects verifying the rule:

$$Confidence(X \rightarrow Y) = \frac{n_{XY}}{n_X}$$

Lift (relative probability) is the ratio of observed support to that expected if X and Y were independent. *N* represents the total number of subjects. It measures the performance of a rule to identify a subgroup from a larger population:

$$Lift\ (X \rightarrow Y) = \frac{N \times n_{XY}}{n_X \times n_y}$$

SIGNIFICANCE TESTS

P-values from two statistical tests are calculated for each association rule:

- A Fisher's exact test, suitable for small sample sizes, is calculated to determine the difference in discretized groups of the antecedent (x) and consequent (y) of an association rule. Since an association rule is directional (i.e. positive relationship between two groups), a right-sided p-value is computed[12].
- A Mann-Whitney U test is also computed to assess the relationship between continuous (outcome) and categorical variables. In this case, a two-sided p-value is calculated in order to be able to evaluate both positive and negative association rules.

These five metrics could be completed by other measures of interestingness such as *leverage*, *gain* and other further metrics currently being developed and discussed in the literature.

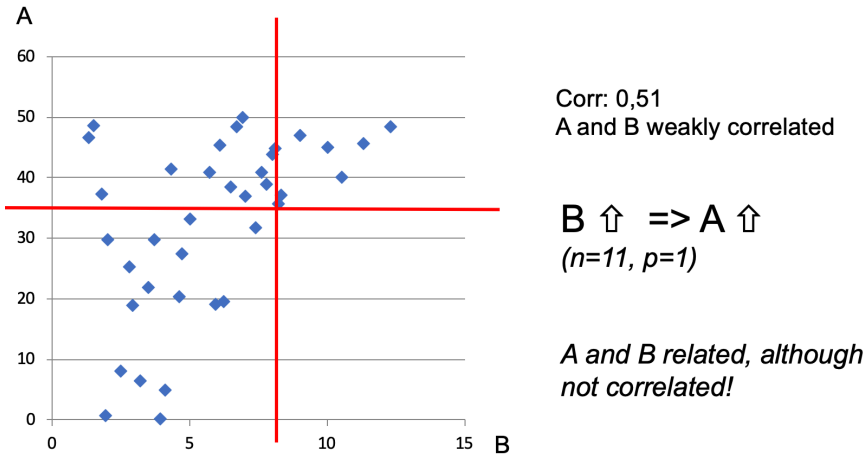
ASSOCIATION RULES

KEM® uses FCA to extract exhaustively all associations between data in the form of uni-directional implications. All attributes (descriptors) in the dataset are first transformed (or discretized) into discrete variables. The discretization is based on either the data distribution or a semantic content (i.e. the thresholds are medically meaningful and actionable). Multiple binning schemes for each variable are considered.

In many cases, relations between variables often only exist within particular, specific ranges rather than globally. These relationships are often dismissed as weak correlations by more traditional approaches when, in fact, the relationships are both relevant and meaningful.

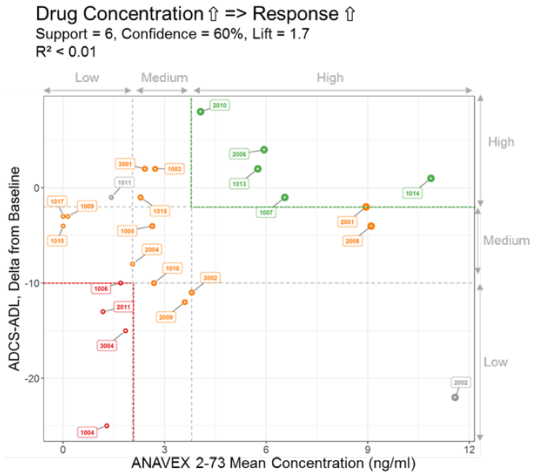
Figure 3 shows a weak correlation between A and B (R2 of 0.51), whereas association rules analysis identifies a strong, one-way relation that links high values (tertiles) of B with A: if B_high then A_high. This relation is uni-directional since the opposite relation if A_high then B_high is not true.

Figure 3. Correlation versus directional non-linear implication.



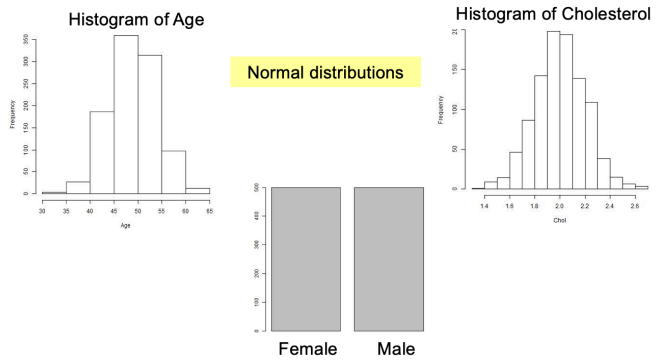
Real example exploring the relation between the concentration of a drug used in Alzheimer's therapy (ANAVEX2-73) and the change of a cognitive score (ADCS-ADL Activity of Daily Living score Delta from Baseline) since beginning of treatment is shown in Figure 4. Patients are represented by circles colored based on outcome higher values i.e. showing lower decline, are desirable (red). We can see that for lower concentration of the drug, the response in spread. As the concentration increases, more patients are responders (red). The relationship presents a weak correlation (R2 < 0.01) however FCA association rule-based analysis identifies a strong relationship of cognitive response for patients having higher concentrations of ANAVEX 2-73`

Figure 4. Relation between the concentration of a drug used in Alzheimer's therapy (ANAVEX2-73) and the change of a cognitive score (ADCS-ADL Activity of Daily Living score Delta from Baseline) since beginning of treatment[13].



An important aspect of the non-supervised analysis is the ability to identify strong links between all parameters and not only between parameters and outcome. Figure 5 represents the distribution of 3 parameters, namely age, cholesterol levels and gender, extracted from the summary report of a 1000 patient clinical trial. The histograms show a normal distribution for all 3 parameters.

Figure 5. Histogram representation of the distribution of 3 parameters from a clinical trial dataset of 1000 patients



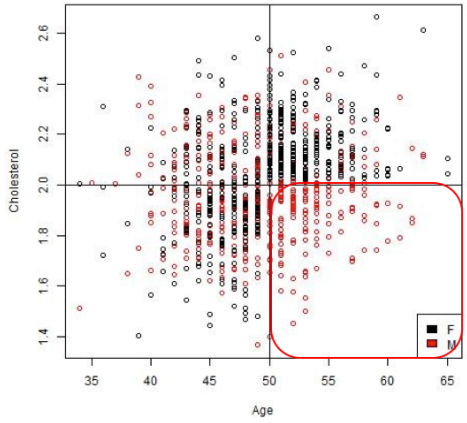
However, a non-supervised analysis identifies that, in this cohort, low cholesterol levels in patients above the age of 50 is only observed in males (Table 1). This relation, shown on the third row of Table 1, is observed in 219 patients (i.e. near ¼ of the dataset) and its confidence is 100% (i.e. it is always true).

Table 1. Extracted association rules

KEM® derived rules	Confidence	Support
Age > 50 => Chol > 2	0.71	301
Gender = F => Chol > 2	0.70	349
(Chol <2) & (Age > 50) => Gender = M	**1.00**	**219**

Representing Age versus Cholesterol levels and using gender to color the data points illustrates this finding as the lower right quadrant contains only male patients. The above identified relation is obvious when visually inspecting this plot (Figure 6). However, the AI system helps perform this type of analysis in a systematic way, without the need of plotting every combination of descriptors. Note that the system does not tell anything about the relevance of the finding that needs to be interpreted in the light of the context in which the data was generated or collected.

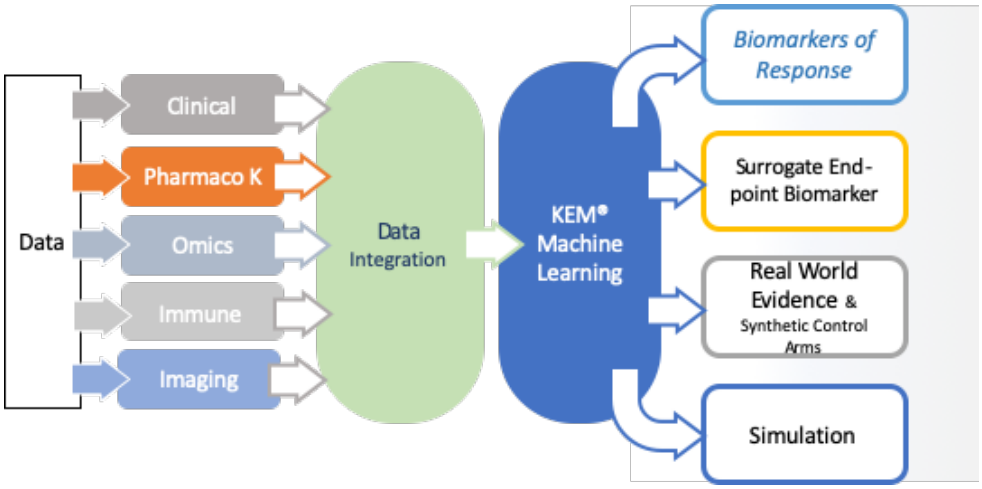
Figure 6. Scatter plot of Cholesterol vs Age and gender (red: male)



THE KEM® KNOWLEDGE EXTRACTION AND MANAGEMENT ENVIRONMENT

The KEM® system has been developed using Java as multiple nodes. The nodes can be used through workflows. A number of workflow systems are supported including KNIME®. The overall architecture of the system is represented in the Figure 7.

Figure 7. The overall KEM® architecture



The current implementation has been tested with datasets as small as 10 patients and up to 134 million patient records. Typical data types include clinical evaluations and questionnaires, blood biochemistry, DNA from whole exomes sequencing and RNA expression, proteomic data from Mass spectrometry and NMR spectroscopy, gut microbiota meta-genomics, pharmacokinetic and Fluorescence-activated cell sorting (FACS) flow cytometry data.

GOING BEYOND BINARY READOUT IN CLINICAL TRIALS

The prospect of a successful clinical trial is notoriously low[14] and a number of factors have been identified as the potential cause of this globally low yield. Patient heterogeneity is high on the list of potential culprits. In oncology, the approach has been to run multiple small, open label clinical trials where a broad range of biomarker candidates are identified. Biomarker and patient inclusion/exclusion hypotheses are generated and tested iteratively, serving as foundation for the design of follow up controlled studies. Biomarker driven clinical trials are thought to have an overall higher success rate[15] (e.g., Crizotinib[16]).

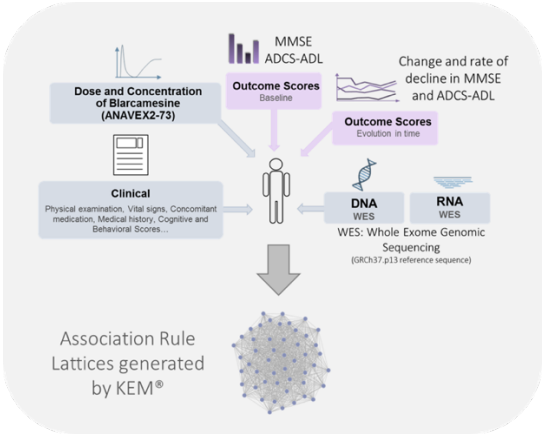
Crizotinib is a good example of accelerated US FDA approval in oncology following an open label Phase 1a study, where a novel biomarker (ALK+) was first identified based on 2 of 11 patients with non-small cell lung cancer enrolled in a Phase 1 dose escalation trial[16]. The biomarker was validated through an amended Phase 2a study including 19 patients, with 10 out of 19 patients being labeled as responders, which ultimately led to a preliminary drug registration. Subsequently, a confirmatory trial enrolled 82 ALK+ patients and showed 57% partial response and 1% complete response[16].
Unbiased and unsupervised data analytics methods such as the KEM® platform may be ideal tools to streamline this type of discovery process. As such, they provide systematic iterative precision medicine tools for drug development, suitable as much to AD as to oncology.

We provide here two examples outside of the oncology space to illustrate our AI software and process, first with an example from clinical trials testing the safety and efficacy of a drug against Alzheimer's disease and second an example in Type 1 Diabetes.

IDENTIFICATION OF GENOMIC BIOMARKERS OF RESPONSE IN ALZHEIMER'S DISEASE CLINICAL STUDY ENABLING A PRECISION MEDICINE APPROACH

Alzheimer's Disease (AD) is an irreversible, progressive and multifactorial neurodegenerative disorder and has no disease modifying treatment. Today it remains an international health priority. Blarcamesine, a selective sigma-1 receptor (SIGMAR1) agonist was investigated in a 57-week Phase 2a study with 32 mild-to-moderate AD dementia patients showing a favorable safety profile. Figure 8 shows the summary of integrated data sources of the ANAVEX2-73-002/003 trial making up a total of 1,152 descriptors. Patient descriptors are shown in grey and outcome data in purple. The integrated data was analyzed using our unsupervised FCA rule-based methodology.

Figure 8. Summary of integrated data sources.

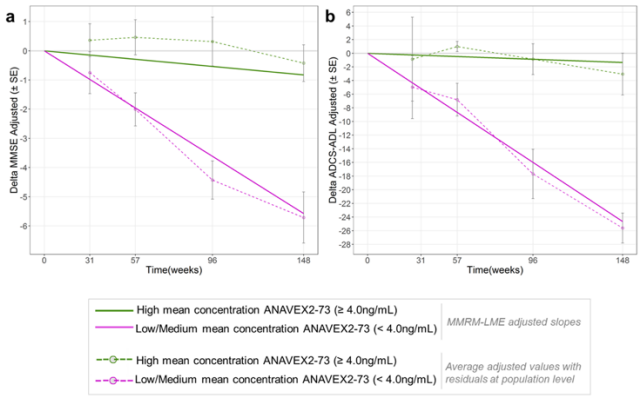


Associations of descriptors linked to outcome at week 57 were identified. Drug concentration, clinical evaluation at baseline and 2 genomic alterations were identified as potential biomarkers of response.

Biomarkers with a significant impact on clinical outcomes were identified at week 57: mean plasma concentration of Blarcamesine (Slope MMSE:p<0.041), genomic variants SIGMAR1 p.Gln2Pro (ΔMMSE:p<0.039; ΔADCS-ADL:p<0.063) and COMT p.Leu146fs (ΔMMSE:p<0.039; ΔADCS-ADL:p<0.063), and baseline MMSE score (Slope MMSE:p<0.015).

The biomarker hypothesis identified at week 57 was validated at week 148 where their combined impact on drug response was confirmed at week 148 with linear mixed effect models (Figure 9) [13].

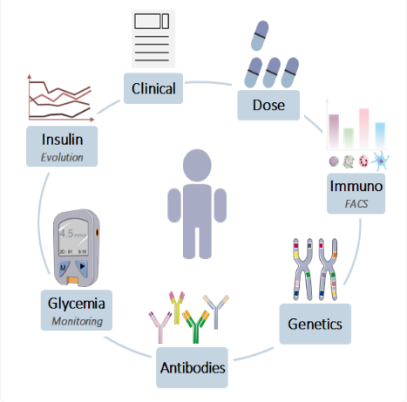
Figure 9. Biomarkers combined impact on drug response was confirmed at week 148 with linear mixed effect models.



TOWARDS A FIRST PRECISION MEDICINE IMMUNOTHERAPY FOR THE TREATMENT OF TYPE 1 DIABETES USING AI

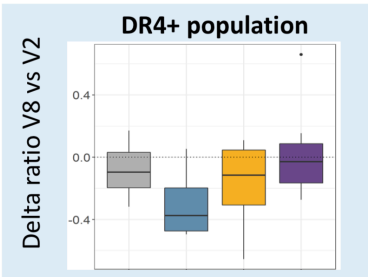
The incidence of type 1 diabetes is rapidly increasing in high-income countries at an annual rate of about 3 per cent, with the disease increasingly occurring in younger children. It affects more than 40 million people worldwide. Currently, the only available treatment is to control the blood glucose level with multiple daily insulin injections. We have been collaborating with a group that is developing active and specific immunotherapeutics (IMCY-0098) for the treatment and prevention of type 1 diabetes ('T1D'), for patients with recent-onset T1D[17]. A total of 41 patients were included and a broad range of parameters were recorded for the patients over a one-year period (Figure 10).

Figure 10. Summary of integrated data sources.



The systematic analysis of the data using our FCA was able to identify a potential biomarker of response where patients expressing HLA-DR4 (DR4+ or DR4 only) and treated with intermediate or high dose show a positive trend for multiple endpoints (Figure 11).

Figure 11. Evolution of C-peptide (AUC MMTT) are represented. Boxplots correspond to Placebo, Dose 1, 2 and 3



These results pave the way for further clinical development of IMCY-0098 in type 1 diabetes and other Imotopes™ for diseases with high unmet clinical need.

CONCLUSION

The proposed analytical strategy can be generalized to all therapeutic areas where patient heterogeneity is an issue. We would advocate that this provides the basis of running smaller and less costly iterative clinical trials that are more effective clinically and ethically.

REFERENCES

1. Hekler EB, Klasnja P, Chevance G, Golaszewski NM, Lewis D, Sim I. Why we need a small data paradigm. *BMC Med*. 2019;17(1):133. doi:10.1186/s12916-019-1366-x

2. Car J, Sheikh A, Wicks P, Williams MS. Beyond the hype of big data and artificial intelligence: building foundations for knowledge and wisdom. *BMC Med*. 2019;17(1):143. doi:10.1186/s12916-019-1382-x

3. Amy Abernethy. *Nat Rev Drug Discov*. 2018;17(7):462-464. doi:10.1038/nrd.2018.101

4. Tran BX, Vu GT, Ha GH, et al. Global Evolution of Research in Artificial Intelligence in Health and Medicine: A Bibliometric Study. *J Clin Med*. 2019;8(3):360. doi:10.3390/jcm8030360

5. Arrieta AB, Díaz-Rodríguez N, Del Ser J, et al. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. *ArXiv191010045 Cs*. December 2019. http://arxiv.org/abs/1910.10045. Accessed February 25, 2020.

6. Doshi-Velez F, Kim B. Towards A Rigorous Science of Interpretable Machine Learning. *ArXiv170208608 Cs Stat*. March 2017. http://arxiv.org/abs/1702.08608. Accessed February 25, 2020.

7. Explainable Artificial Intelligence. https://www.darpa.mil/program/explainable-artificial-intelligence. Accessed February 26, 2020.

8. Jullian N, Jourdan N, Afshar M. Hypothesis Generation for Scientific Discovery. Examples from the Use of KEM®, a Rule-Based Method for Multi-Objective Analysis and Optimization. *IOS Press*. 2008;Volume 9 : Towards Drugs of the Future: Key Issues in Lead Finding and Lead Optimization:75-80.

9. Lee H, Yune S, Mansouri M, et al. An explainable deep-learning algorithm for the detection of acute intracranial haemorrhage from small datasets. *Nat Biomed Eng*. 2019;3(3):173-182. doi:10.1038/s41551-018-0324-9

10. Agrawal R, Imieliński T, Swami A. Mining Association Rules between Sets of Items in Large Databases. In: New York, NY, USA; 1993:207-216. http://dx.doi.org/10.1145/170036.170072. Accessed November 12, 2018.

12. Hahsler M, Hornik K. New Probabilistic Interest Measures for Association Rules. *Intell Data Anal*. 2007;11(5):437–455.

13. Hampel H, Afshar M, Parmentier F, et al. Full Genomic Analysis of ANAVEX®2-73 Phase 2a Alzheimer's Disease Study Identifies Biomarkers Enabling Targeted Therapy and a Precision Medicine Approach. *Alzheimers Dement J Alzheimers Assoc*. 2018;14(7):P1519-P1520. doi:10.1016/j.jalz.2018.07.027

14. Hay M, Thomas DW, Craighead JL, Economides C, Rosenthal J. Clinical development success rates for investigational drugs. *Nat Biotechnol*. 2014;32(1):40-51. doi:10.1038/nbt.2786

15. Wong CH, Siah KW, Lo AW. Estimation of clinical trial success rates and related parameters. *Biostatistics*. doi:10.1093/biostatistics/kxx069

16. Ou S-HI. Crizotinib: a novel and first-in-class multitargeted tyrosine kinase inhibitor for the treatment of anaplastic lymphoma kinase rearranged non-small cell lung cancer and beyond. *Drug Des Devel Ther*. 2011;5:471-485. doi:10.2147/DDDT.S19045

17. Van Rampelbergh et al. Phase Ib clinical trial of IMCY-0098 in young adults with recent-onset type 1 diabetes. 55th Annual Meeting of the European Association for the Study of Diabetes, September 2019, Barcelona, Spain.
.

CONTACT INFORMATION

Your comments and questions are valued and encouraged.

Contact the author at:

Mohammad Afshar
Ariana Pharma
43-Av Grande Armée
75116 Paris
France

m.afshar@arianapharma.com
Web: www.arianapharma.com