# The FDA contribution to Health Data Science

Francesca Ieva

**Abstract** This contribution aims at presenting examples of Health Data Science where advanced methods based on Functional Data Analysis are used to bring value to clinical and biological problems.

**Key words:** Healthcare Research, Health Data Science, Health Analytics, Functional Data Analysis

## 1 General Background and motivations

Healthcare research generates a significant portion of big data from administrative routine, clinical practice, molecular sources, imaging investigations. This data is of paramount interest for defining a comprehensive fingerprint of patients' status to be used in primary and secondary prevention, therapy endpoints predictions and scoring [1]. This requires proper management and analysis in order to derive meaningful information from diversified data sources [2]. There are various challenges associated with each step of handling healthcare data, ranging from data access and integration, to development of advanced models for fingerprint extraction, to issues in late fusion of this information into suitable predictive models [3] [4]. That is why, to provide relevant solutions for improving public health, healthcare providers are required to be fully equipped with appropriate infrastructure to systematically generate and analyze data.

Among others, in this paper we focus on situations where the interest lies in dealing with time-varying processes, i.e., phenomena evolving over time. Examples are the dynamic monitoring of biological or vital signals, or models for longitudinal observations and covariates to be properly summarized and treated for plugging them

Francesca Ieva
MOX lab, Department of Mathematics, Politecnico di Milano, Milan , Italy & Health Data Science Center, Human Technopole, Milan, Italy e-mail: francesca.ieva@polimi.it

into Statistical and Machine Learning models and algorithms. In all these cases, Functional Data Analysis (FDA)[5] may be used as a proficient support to precision medicine, since it allows for developing powerful methods which account not only for baseline or cross sectional information, but also for the dynamic of the process at hand.

In particular, in Sect. 2 an overview of clinical applications where models exploiting FDA techniques are used will be presented, with the aim of highlighting FDA potential in supporting clinical practice and precision medicine approach. The first one (Sect. 2.1) concerns the extraction of dynamic information about patterns of care from Healthcare Utilization Databases. In the second case (Sect. 2.2), the dynamic monitoring of longitudinal biomarkers is presented. Finally, the third application (Sect. 2.3) is related to the assessment of genotype association with specific phenotypes of interest.

## 2 Case Studies

The case studies proposed in this Section are part of current researches carried out within the Health Analytics group at MOX lab (Department of Mathematics, Politecnico di Milano) and at the Health Data Science Center of Human Technopole. In particular, we exploit results coming from [6], [7] and [8].

### *2.1 Functional modeling of recurrent events on time-to-event processes*

In clinical practice, it is often the case where the association between the occurrence of events and time-to-event outcomes is of interest; thus, it can bemodeled within the framework of recurrent events. The purpose of our study is to enrich the information available for modeling survival with relevant dynamic features, properly taking into account their possibly time-varying nature, as well as to provide a new setting for quantifying the association between time-varying processes and time-to-event outcomes.

In [6] and [7] we propose and discuss an innovative methodology to model information carried out by time-varying processes by means of functional data, modeling each time-varying variable as the compensator of marked point process the recurrent events are supposed to derive from. By means of Functional Principal Component Analysis, a suitable dimensional reduction of these objects is carried out in order to plug them into a Cox-type functional regression model for overall survival. We applied our methodology to data retrieved from the administrative databases of Lombardy Region (Italy), related to patients hospitalized for Heart Failure (HF) between 2000 and 2012. We focused on time-varying processes of HF hospitalizations and multiple drugs consumption and we studied how they influence patients' over-

all survival. This novel way to account for timevarying variables allowed to model self-exciting behaviors, for which the occurrence of events in the past increases the probability of a new event, and to quantify the effect of personal behaviors and therapeutic patterns on survival, giving new insights into the direction of personalized treatment.

## *2.2 A wavelet-mixed landmark survival model for the effect of short-term oscillations in longitudinal biomarker's profiles*

In many chronic diseases, patient's disease progression and status can be monitored over time through easily measured biomarkers. Medical decisions regarding treatments are often made on the basis on such monitoring. Hence, it is important to have quantitative tools to exploit information given by such measurements. Since the final goal of medical decisions is the minimization of the risk of adverse events such as hospitalizations or death, survival models are very often the building blocks of such tools.

Recently two methods have been in widespread use for the modeling of longitudinal internal time-dependent covariates and survival: joint models and landmark models [9]. In joint models the longitudinal biomarker process is modeled through linear mixed effects models which allow to consider the subjects' specific trajectories of the biomarker through the inclusion of random effects into the model. These latent variables are used to model the effect of unobserved variables that are responsible of subjects' deviation from the overall mean trajectory specified through the fixed effects [10].

One of the main advantages of modelling the longitudinal process of the biomarker is that we can study the relationship between the rate of change of the biomarker and

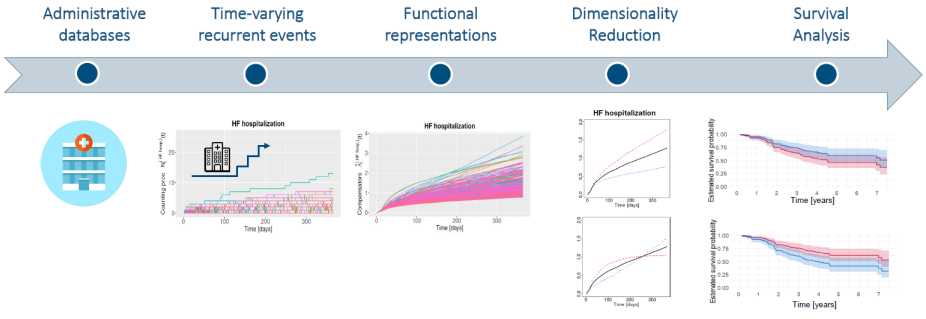


**Fig. 1** Example of workflow for Healthcare Utilization Databases (HUD) exploitation. Functional data here represent the compensators of suitable stochastic processes describing re-hospitalizations over time and drug purchases for a given diseases of interest. They are analytically treated to be properly plugged into a predictive model for the main endpoint of the study (here long-term any cause survival).

the time-to-event process. However, a limitation of existing methods is that they don't take into account that the time scale of the survival process is very different from the time scale at which changes in the biomarkers happen in the human body. While in observational studies with the survival as the outcome of interest the follow-up period can be very long (e.g. years), intervals of measurements times are highly irregular because they depend of the clinical requirements. For example, during drugs up-tritation or acute disease periods more frequent measurements are required. As a consequence, such data is characterized by a very high observation period/time-between-measurements ratio. Sudden changes in biomarkers are very often important from a prognostic point of view. However, linear mixed effects model consider transient changes as measurement error when the length of the follow-up is much greater then their duration. Therefore, both joint models and landmark mixed models implicitly make a strong assumption, i.e., that short-term oscillations in biomarkers are either not present or they don't have effect on the risk of adverse events.

In [8] we propose a novel approach to study the association between a continuous timedependent longitudinal covariate and a time-to event outcome in order to overcome this limitation. Our method allows to identify and study the role of the short-term oscillations of the biomarker over time via a wavelet based functional approach and to set up a dynamic monitoring tool to support clinical decision making. The method is based on coupling a linear mixed effect model with a wavelet filter. The first allows to identify the effect of fixed covariates and the long-term effect of time. On the other hand, the wavelet filter is used to identify the subject-specific shortterm oscillations. The main idea is to combine a between-subjects model with a within-subject method such as a wavelet transform to obtain a functional and subject-specific representation of the biomarker trajectory over time [11]. Moreover, FDA offers methods to obtain functional objects from discrete and noisy longitudinal data and its application on time-dependent biomarker covariates offers a novel approach to extract biomarkers behaviors over time not observable with other methods.

## *2.3 Genomic trajectories*

It is often the case in healthcare research that not the single measurement related to a quantity of interest is informative of the patient's status, but the evolution (a.k.a. behavioral pattern) of this quantity over time is. This is actually becoming particulary true in genomic studies, where the association between genomic traits and phenotypes are the goal of the analysis [12].

In such cases, the development and application of novel methodologies to study the genetic architecture of biomarkers' evolution over time, and their relationship with clinical outcomes of interest is the focus, and needs FDA application for both representing longitudinal trajectories and then modeling their effect on suitable endpoints as well as to assess their association with other clinical and biological traits.

Primary care and hospital administrative data can be exploited to derive longitudinal biomarker trajectories. Modelling Gene-biomarkers associations (considering single or multiple trajectories at a time) will result in the identification of representative biomarkers trajectory groups, and their association with individuals' genetic background. The considered biomarkers may include BMI, cholesterol, blood pressure, and others. Further clinical time varying information can be retrieved and included in modelling gene-biomarker relationships, to adjust for exogenous confounders influencing biomarkers' evolution, such as prescriptions and health conditions.

Once the trajectories are pointed out, further analyses will identify the effect of specific trajectory groups on the clinical endpoints (for the case of interest, ischemic stroke, coronary artery disease and diabetes). This enables novel insights on how biomarkers evolution over life time can increase the risk of adverse events and/or diagnoses. The application is carried out on UK Biobank data [13].

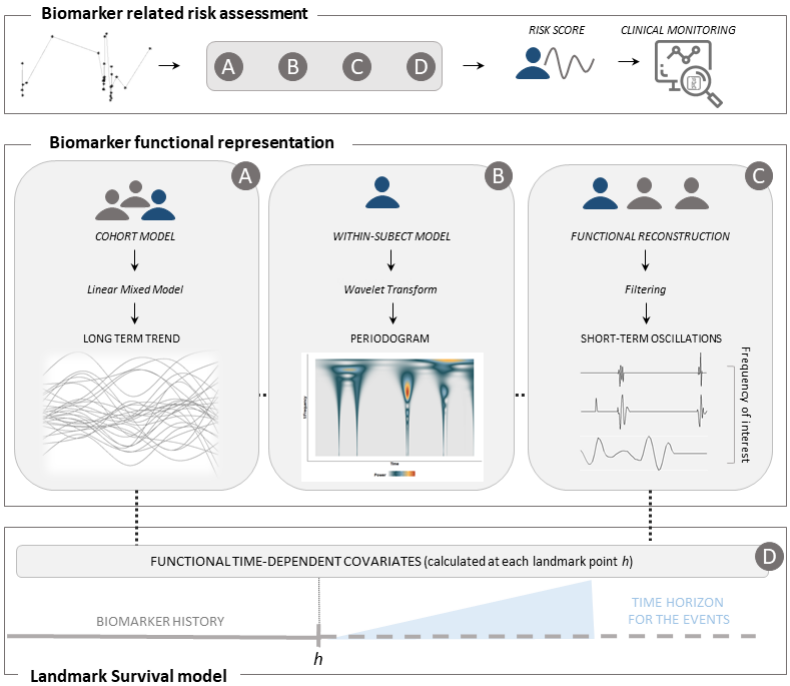


**Fig. 2** Example of workflow for biomarkers monitoring in the context of Diskalaemya for Heart Failure patients.

## 3 Conclusions

In this paper we tried to emphatize the role FDA may have in different part of the health analytics process. The combination of advanced statistical methodologies with challenges proposed by recent clinical and biological problems may end up in a re-shaping of future directions of research in healthcare setting.

**Acknowledgements** This work was possible thanks to the substantial contribution of Dr. Caterina Gregorio, Michela Massi, Marta Spreafico and all the people dedicating their passion and capabilities to the healthcare research programs within the Health Analytics unit at Department of Mathematics of Politecnico di Milano and the Health Data Science Center of Human Technopole.

## References

1. Schüssler-Fiorenza Rose SM et al: A longitudinal big data approach for precision health. Nat. Med. **25**, 792-–804 (2019)
2. Amal S, Safarnejad L, Omiye J A et al.: Use of Multi-Modal Data and Machine Learning to Improve Cardiovascular Disease Care. Frontiers in Cardiovascular Medicine, **9** (2022)
3. Stahlschmidt SR, Ulfenborg B, Synnergren J: Multimodal deep learning for biomedical data fusion: a review, Briefings in Bioinformatics, **23**(2): bbab569 (2022)
4. Acosta JN, Falcone GJ, Rajpurkar P et al. Multimodal biomedical AI. Nat Med 28, 1773–1784 (2022)
5. Ramsay JO, Silvermann BW: Functional Data Analysis. Springer-Verlag New York (2005)
6. Spreafico M, Ieva F: Functional modelling of recurrent events on time-to-event processes. Biom. J., **63**(5): 948-–967 (2021)
7. Spreafico M, Ieva F: Dynamic monitoring of the effects of adherence to medication on survival in Heart Failure patients: a joint modelling approach exploiting time-varying covariates. Biom. J., **63**(2) Special Issue: Novel Aspects in Biostatistics: 305–322 (2021)
8. Gregorio C, Barbati G, Ieva F: A wavelet-mixed effect landmark model for the effect of short-term oscillations in longitudinal biomarker's profiles on the risk of death: an application for the monitoring of potassium in Heart Failure. ArXiv preprint (2022) https://doi.org/10.48550/arXiv.2204.05870
9. Rizopoulos D, Molenberghs G, Lesaffre E M.E.H.: Dynamic predictions with time-dependent covariates in survival analysis using joint modeling and landmarking. Biom. J. **59**(6), 1261-–1276 (2017)
10. Tsiasis A, Davidian M: Joint modeling of longitudinal and time-to-event data: an overview on JSTOR. Stat. Sin. **14**(3), 809—834 (2014)
11. Unser M, Aldroubi A: A review of wavelets in biomedical applications. Proceedings of the IEEE **84**(4), 626—638 (1996)
12. Auton A et al: A global reference for human genetic variation. Nature **526**(7571), 68—74 (2015)
13. Sudlow C et al.: UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. PLoS Med **12**, e1001779 (2015)