Research Proposal

ALL-EMBRACED aims to offer the following innovative services: 1) pre-screening of patients representative of the general population by exploiting natural language-based systems; 2) generation of auxiliary synthetic data (via Synthea) to increase simulations and determine the best possible eligibility criteria; 3) relaxation of inclusion/exclusion criteria through an innovative approach (i.e.: SHAPLEY). All the previous services materialize in 4) optimized plannings of CTs.
An analysis of the effects of different eligibility criteria on clinical outcomes will be performed based on large sets of retrospective CTs. This will allow to define new guidelines for future CTs aimed at different pathologies, increasing the number of patients and validating, equally, the safety aspect. The ultimate goal is to trigger a paradigm shift in which CTs will be more inclusive with optimized planning and selection of patients.

Research Methods

The work plan of ALL-EMBRACED project brings together the technology developed by the IRCCS-FBF within the framework of web-platforms [7] and explainable AI approaches [8] from a Technology Readiness Level (TRL) 4 (technology validated in a lab) to at least TRL 7 (system prototype demonstration in operational environment), providing a working demonstrator that can be packaged into a product that can be used in clinical and research environments.
It will advance the NLP approaches developed in IRCCS-FBF [9] from TRL 3 (experimental proof of concept) to TRL 6 (technology demonstrated in relevant environment), with the path to a later TRL being predetermined by experience with the current technology. We will achieve this with a two-year project divided into the following WPs:
WP1: Web-portal creation. The goal of this work-package is to create the free of charge ALL-EMBRACED web-platform as “Software as a Service” (SaaS). We will exploit a LAMP (Linux, Apache, MySQL, PHP/Perl/Python) bundle. The web-portal will be created with the WordPress Content Management System (CMS).
WP2: Data integration. In this work package, we will gather and prepare the data to be used in the ALL-EMBRACED project. In order to assess the representativeness of a CT target population, we must first compare it with the micro-data of the general population. In doing so, we rely the work of the IRCCS-FBF (5 interventional and 17 observational studies) and other CT providers (e.g.: YODA, AACT and CCTI database from clinicaltrials.gov). In addition to CT real-world data, the synthetic data generator service will be developed to further validate the relaxation of eligibility criteria. The CT-data gathered will be used just to define the best eligibility criteria and in no way will be redistributed via ALL-EMBRACED.
WP3: Relaxation of eligibility criteria to increase representativeness in CT studies. This WP implements and tests a novel approach to evaluate the influence of individual eligibility criteria in several similar CT studies. We hypothesize that our approach will delineate a new landscape to conduct CT. We will demonstrate our framework comparing the same clinical outcomes of the original CTs versus the same clinical studies where relaxation of the eligibility criteria will be simulated.
WP4: ALL-EMBRACED prototype. In this WP, the final prototype interface will be launched.

Plan to use YODA individual data with synthetic data

Most CT databases (such as clinicaltrials.gov, ISRCTN, and others) typically provide aggregated information on biomarkers and clinical outcomes. However, disaggregated data (i.e. individual data

for a large number of patients) would be extremely useful to assess the distribution and treatment trends for different NCD pathologies. A tool to generate reliable synthetic data at the patient-level would address the lack of real-world patient-level data.
In the ALL-EMBRACED project, the generation of synthetic data is done through Synthea (https://github.com/synthetichealth/synthea), an open source tool that can generate the medical history for synthetic subjects, based on results from academic publications and statistical institutions, reproducing the prevalence of different pathologies. Of all the medical data, Synthea can reproduce the similar physical and neurological measurements collected as biomarkers in CTs. However, these measurements need to be validated using real world CT patient measurements.
Individual data from the YODA project will not be used to generate synthetic data in Synthea, but will only serve as comparative data to check the reliability of the Synthea data. The Mini Mental State Examination (MMSE), for example, is a standard neuropsychological tests, which is collected for patients with Alzheimer's Disease. The results of this test are both produced by Synthea for AD patients and provided under the YODA individual data. The average and the quartiles for both groups are calculated and compared. If there are no statistical differences between the two sets, the MMSE generation process is validated by Synthea. This validation procedure will be repeated for each biomarker of interest available in different CTs, proven to be relevant in monitoring the pathologies of interest in ALL-EMBRACED.

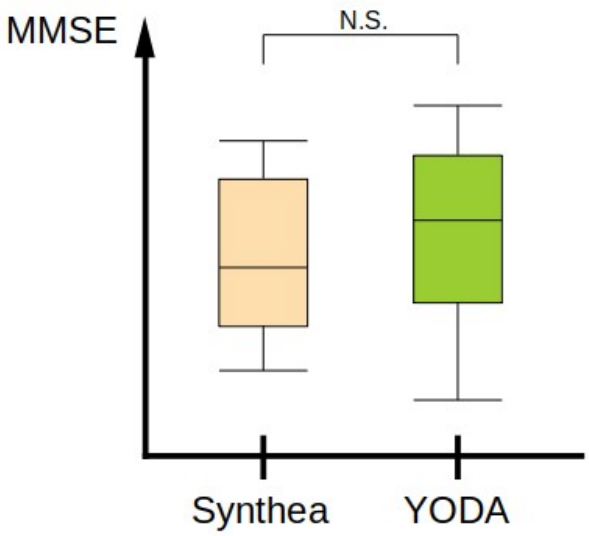

Figure: Synthea is expected to generate and mimic real CT studies without statistical differences (Alpha = 0.05, p-value = not statistically different (N.S.))