Borsa di studio attivata ai sensi di quanto disposto dal D.M. n. 1061 del 10/08/2021

Titolo del progetto: Covid-19 patients and HCW cohorts data harmonisation, standardisation and processing for analysis of data driven evidence of SARS-CoV-2 variants impact

La borsa sarà attivata sul seguente corso di dottorato accreditato per il XXXVII ciclo:
DATA SCIENCE

Responsabile scientifico: Stefano Leonardi
Area per la quale si presenta la richiesta: INNOVAZIONE
Numero di mensilità da svolgere in azienda: 6
Azienda: Euresist Network
Il Dipartimento è disponibile a cofinanziare per un importo pari a euro: 7000
Dipartimento finanziatore: DIPARTIMENTO DI INGEGNERIA INFORMATICA, AUTOMATICA E GESTIONALE - ANTONIO RUBERTI- con delibera del 20 Settembre 2021

Progetto di ricerca:
Il progetto di dottorato mira a sviluppare metodi di data science per l'integrazione dei reali per il monitoraggio della pandemia COVID-19 e delle sue varianti provenienti da 18 centri in quattro continenti che collaborano al progetto UE EUcare guidato da Euresist Network che è stato finanziato dall'UE per lo studio delle varianti SARS-CoV-2. L'obiettivo finale è fornire prove solide e basate sui dati per affrontare le varianti di SARS-CoV-2 e le epidemie di COVID-19.
I centri in Europa, Asia, Africa e America Centrale forniranno una grande varietà di vaccini, dati sanitari, distribuzioni di varianti virali e protocolli di trattamento. Verranno istituite una coorte di pazienti, una coorte di pazienti "covid-19 lungo" e una coorte di operatori sanitari con raccolta di dati prospettici e retrospettivi

L'enorme quantità di informazioni retrospettive già disponibili e che verranno raccolte in prospettiva sarà integrata con dati provenienti da altri studi in corso e da raccolte di dati aperti disponibili, per generare un insieme di dati enorme e di grande dimensione.
Tuttavia, tutte queste diverse fonti di dati possiedono una grande varietà di formati di dati, codifiche e procedure di raccolta che ne ostacolano l'usabilità.
Le specifiche dei dati dovranno essere armonizzate in collaborazione con i medici coinvolti e tenendo conto di studi e iniziative pertinenti in corso riguardanti le specifiche dei dati COVID-19 (ad esempio il progetto RECODID Hadea e le indicazioni dell'OMS). Il progetto fornirà l'armonizzazione delle definizioni e delle misure di prestazione per selezionare un insieme minimo di variabili di consenso "fondamentali". Inoltre, sarà sviluppata una serie di procedure operative standard (SOP) per standardizzare gli approcci tra i centri alla memorizzazione di dati e campioni. Queste SOP per le specifiche dei dati (che incorporano le variabili dei dati di consenso di base e quelle "opzionali") saranno disponibili gratuitamente per studi di ricerca in tutto il mondo, con revisioni/aggiornamenti regolari, se necessario.
Il progetto di dottorato è quindi finalizzato a progettare la pipeline di data science per l'acquisizione, la pulizia e l'integrazione in un database centralizzato o federato dei dati COVID-19 e delle sue varianti. I dati estratti da ciascun centro nello studio saranno verificati per coerenza e completezza mediante diversi metodi di data mining, con rapporti sulla qualità dei dati e query generate per ciascuna coorte.
I dati saranno apertamente accessibili e riutilizzabili in conformità con i principi FAIR.

Titolo del progetto (inglese): Covid-19 patients and HCW cohorts data harmonisation, standardisation and processing for analysis of data driven evidence of SARS-CoV-2 variants impact

Progetto di ricerca (inglese):
The phd project aims at developing data science methods for the integration of real life data for monitoring of the COVID-19 pandemic and its variants from 18 centres in four continents that collaborate in the EU project EUcare led by Euresist Network GEIE that has been funded by EU to study SARS-CoV-2 variants. The final goal is to provide robust, data driven evidence to deal with SARS-CoV-2 variants and COVID-19 epidemics.
Centres in Europe, Asia, Africa and Central America will provide a large variety of vaccines, health care systems, viral variants distributions and treatment protocols. A patients' cohort, a "long COVID-19" patients' cohort and an healthcare workers cohort will be set-up with collection of prospective as well as retrospective data
Table 1. Retrospective available information and prospective enrolment estimates.

The huge amount of retrospective information already available and that being collected prospectively will be complemented with data from other ongoing studies and from available open data collections, to generate a huge, highly dimensional data set.
However, all these diverse data sources come with a big variety of data formats, coding and collection procedures which hampers its usability.
Data specifications will need to be harmonised in collaboration with the involved clinicians and taking into account ongoing relevant studies and initiatives regarding COVID-19 data specifications (e.g. the RECODID Hadea project and the WHO indications). The project will provide harmonisation of definitions and outcome measures for a minimal set of "core" consensus variables. Also, a range of standard operating procedures (SOPs) will be developed to standardise approaches across the centres relating to data and samples. These SOPs for data specifications (incorporating the core consensus data variables as well as "optional" ones) will be freely available for research studies worldwide, with regular reviews/updates as required.
The PhD project is therefore aimed to design the data science pipeline for the acquisition, cleaning and integration into a centralised or federate database of the COVID-19 data and its variants. Data extracted by each centre in the study will be checked for consistency and completeness by several data mining methods, with data quality reports and queries generated for each cohort.
Data will be openly accessible and reusable in accordance with the FAIR principles.