DATA SCIENCE FOR HEALTH

Studying the representativeness of clinical trials to develop more effective treatments for all

Call deadline

January 30th, 2024

CALLS 2023

Fondazione CARIPLO

TUTE SERVARE MUNIFICE DONARE • 1816

In collaboration with

Index

1. INTRODUCTION

"Data Science for Health - Studying the representativeness of clinical trials to develop more effective treatments for all" is a call promoted by the Scientific Research Area falling in Fondazione Cariplo strategic goal "Scientific research: supporting multidisciplinary research for the well-being of people and the socio-economic development of communities".

The call is in collaboration with Novartis Farma S.p.A., which provides the dataset detailed in Appendix for public interest and research purposes.

2. BACKGROUND

The health and well-being of the population are invaluable goods and clinical research is a powerful tool to support them. Clinical trials, in fact, aim at obtaining real evidence regarding efficacy and safety of new therapies or diagnostic procedures, with the final goal of improving individual and collective health. Although in principle participation in clinical studies is open to everyone, compliance with inclusion and exclusion criteria, whose planning and evaluation is complex, is mandatory. As recently stated by a publication in Nature journal, excessively restrictive eligibility criteria can represent a serious drawback for the validity of the study itself[1]. Moreover, the U.S. Food and Drug Administration points out that some populations are usually excluded or underrepresented in clinical trials. Furthermore, patients' self-selection - for cultural, social, logistical or time availability reasons - is not unusual. As a consequence, efficacy and safety are sometimes tested on subjects that do not fully represent the overall population who could potentially benefit from that medication and who will actually take it after its approval[2]. Another recent publication highlights how inclusiveness and diversity are important challenges in clinical research: stratification facilitates the detection of significant effects, clarifies the observed variability, reduces confounders and promotes reproducibility[3].

3. AIMS

In light of the scenario outlined above, the call intends to support research projects in the field of Data Science that analyze the representativeness gap between the populations enrolled in clinical trials and the general population affected by the same pathology.

Starting from the dataset provided by Novartis Farma S.p.A. and detailed in Appendix, applicants are required to formulate original working hypotheses, accurately describe the techniques and methods of Data Science that they intend to apply and develop models and guidelines to foster more effective treatments for all. Should additional datasets be used for comparative analysis, applicants will take care to describe them in detail, specifying their added value for the achievement of the project objectives.

Where proposals are submitted in partnership, it will be essential to clarify the expertise contributed by each project partner and how this complements that of the leading organization in a unified design.

In order to stimulate the public debate and promote the adoption of models and guidelines developed, applicants should identify specific ways of sharing the evidence obtained as a result of the research, produce position papers and policy briefs

[1] Evaluating eligibility criteria of oncology trials using real-world data and AI. Nature, 592:629–633 (2021).

[2] Enhancing the Diversity of Clinical Trial Populations - Eligibility Criteria, Enrollment Practices, and Trial Designs Guidance for Industry (https://www.fda.gov/media/127712/download).

[3] Inclusion and diversity in clinical trials: Actionable steps to drive lasting change. Contemp Clin Trials, 116:106740 (2022).

and engage the relevant stakeholders in line with the principles of Responsible Research and Innovation (RRI)[4].

Finally, given the labour market needs for professional profiles and skills in the field of Data Science, proposals shall offer training opportunities and professional growth to the researchers involved.

4. GUIDELINES

4.1 Eligible subjects

Public or private non-profit organizations conducting scientific research. Third sector organizations (e.g.: patients associations) can participate as partners only.

It is mandatory for the operational headquarter of the leading organization to be located within the geographical area of intervention of Fondazione Cariplo (Lombardy and provinces of Novara and Verbano-Cusio-Ossola). This requirement does not apply to potential partners.

By participating to the call, organizations allow the portability of the grant. The transfer of research and funds to another organization will be evaluated by Fondazione Cariplo taking into account the provisions of the present call, the guidelines for submission of proposals ("Guida alla presentazione dei progetti su bandi") and the general criteria for granting awards ("Criteri generali per la concessione dei contributi"), available on Fondazione Cariplo website.

4.2 Eligible projects

To be eligible, projects must meet the following requirements:

- develop independent research for public utility purposes;
- propose a research design using methods and techniques of Data Science;
- use the dataset made available by Novartis Farma S.p.A.;
- comply with the scientific priorities detailed in section 3 "Aims";
- identify a researcher with a maximum age of 45 years, reached by the deadline of the call, as the principal investigator of the leading organization[5].

Principal investigators (PIs) of all the involved organizations (leading and partner) are allowed to:

- submit one proposal only to the call;
- participate as PI in more than one call of the Scientific Research Area, as long as no overlap is displayed between proposals[6];
- not be the PI of any other ongoing proposal previously funded by the Scientific Research Area (except for calls in partnership with other granting bodies)[7].

Proposals involving human subjects, human biological material, human samples and/or personal data must comply with Italian laws.

The total cost of the project should include additional costs[8] and overhead only, according to the eligibility criteria and the thresholds listed below:

- A03 - "Equipment and software"
  This entry shall not exceed 30% of the additional costs. It fully covers the costs for

[4] The European movement on the theme of RRI is constantly evolving, as emerges from the debate which has recently led to the so called "Rome declaration on RRI" (https://ec.europa.eu/research/swafs/pdf/rome_declaration_RRI_final_21_November.pdf). Considering the available literature, it is possible to describe RRI as a dynamic and iterative process which intends to match research and innovation to values, needs and expectations of the society. Moreover, RRI aims at actively involving all the stakeholders taking part to activities of research and innovation, making them mutually responsible with respect to both the research process and its results.

[5] Extension of eligibility:
- maternity leave: 18 months for each child;
- paternity leave: effective period of paternity leave for each child;
- serious illness leave (over 90 days): actual period taken for each illness;

[6] In case of overlapping proposals, only the first one received will be considered.

[7] A proposal is considered ongoing unless the final grant reports - both the scientific and the financial one - have been uploaded on our website before the deadline of this call.

[8] Sum of: A03, A04, A06, A07, A08, A10.

newly acquired equipment and software of multi-year use acquired *ex novo*. It is mandatory to clearly motivate these costs according to the specificity of the project.

- A04 - "Others amortizable costs"
  This entry can include patent costs, as well as costs related to equipment rental.
- A06 - "Temporary staff"
  This entry shall include research personnel costs. The cost for administrative staff is not allowed.
- A07 - "Sub-contractors and consultants"[9]
  This entry can include costs of accessing databases and utilities for the project such as, for example, computing power or virtual machines, both in cloud and remote.
- A08 - "Materials and supplies"
  This entry cannot include costs related to office supplies and photocopies.
- A09 - "Overheads"
  This entry cannot exceed 5% of the additional costs.
- A10 - "Travel, publication and dissemination costs"
  This entry can include travel expenses and conference participation fees for the research personnel involved in the project, costs for meetings between partners, publications costs, training and dissemination activities costs.

The grant will cover 100% of project costs; therefore, the total cost of the project should correspond to the requested amount.

Requested funding must not exceed 200.000 Euros.

Projects must have a maximum duration of 24 months.

Proposals should provide the following mandatory documents:

- Host Institution agreement[10];
- Partnership agreement[11];
- Project form as PDF file[12];
- Detailed budget as EXCEL file[13].

By participating in the call, organizations accept and undertake to respect the indications contained in the following documents, available online for the download: "Policy della Fondazione Cariplo in tema di tutela della proprietà intellettuale", "Policy di open access" e "Linee guida per la citazione del contributo nelle comunicazioni scientifiche". It should be noted that all publications that will derive from the project results will have to report the affiliation of the organization with which the proposal was submitted. In addition, for the purposes of preparing the proposal, it is recommended to carefully read the FAQs specifically prepared for this call.

4.3 Criteria

The evaluation of the received proposals will be carried out according to the methods outlined in the document "Guida alla presentazione dei progetti su bandi"[14] and the criteria represented below:

1. SCIENTIFIC QUALITY (weight 30%)

- clarity in the formulation of objectives and proposed strategies;

[9] This entry should also include audit costs, where applicable. Exclusively certificated auditors are considered eligible. For the sole purpose of ascertaining whether the obligation to perform a financial audit applies or not, the amount which has to be considered is obtained by multiplying the eligible costs detailed in the Budget (sum of entries A06, A07, A08, A09 and A10) by the coefficient 1,25. It is also recommended to refer to the "Grant management and reporting guide" and to the guidelines for audit assignment "Linee guida per affidamento audit"; both documents are available on the Cariplo Foundation's website. However, it is anticipated that in case of funding, the Foundation will confirm to each recipient any obligation to satisfy such fulfillment.

[10] Host Institution agreement should be filled in according to the available online format.

[11] Partnership agreement form should be provided exclusively in case the project involves partner and should be filled in according to the available online format.

[12] The project form should be drawn up on the basis of the form made available for the call.

[13] The detailed budget Excel file will be automatically generated in the "Economic plan" section of the platform.

[14] Since 2020, the Foundation intends to gradually increase its focus on the environmental and climate impacts of funded projects. For more details, see section 10 of the guide.

- relevance of the project with respect to the state of the art;
- presence of a logical and well-structured research design;
- suitability of the proposed methods and techniques with respect to the research objectives;
- originality of the project in terms of the proposed methods and techniques.

2. OUTCOMES (weight 25%)

- expected impact in terms of developing new knowledge;
- expected impact in terms of benefits to citizens and community;
- identification of adequate modalities for sharing project's results and project's ability to foster the public debate on the topic by involving the relevant stakeholders;
- expected impact in terms of project's contribution in the development of new models and guidelines.

3. PRINCIPAL INVESTIGATOR, TEAM AND ORGANIZATIONS (weight 10%)

- PI's scientific and managerial leadership;
- expertise of the research team members;
- project's contribution in terms of scientific and professional growth;
- suitability of infrastructure and tools to the size and type of the project.

4. RELEVANCE FOR THE FOUNDATION (weight 15%)

- expected impact in terms of value creation for the local community.

5. CONSISTENCY WITH THE DATASET (weight 10%)

- compatibility of the research hypothesis and objectives with the features of the dataset provided.

6. BUDGET AND PROJECT DURATION (weight 10%)

- budget suitability, coherence and duration of the project.

Please note that proposals' evaluation is carried out through peer review, meaning a scientific merit assessment entrusted by qualified experts who are requested to follow specific rules and procedures aimed at excluding conflicts of interest, ideological prejudices, personal pressures and self-referentiality. Fondazione Cariplo will express opinions limited to the relevance for the Foundation and the coherence of the budget. Novartis Farma S.p.A. will assess the consistency with the dataset.

4.4 Non-eligible projects

Proposals will be considered non-eligible in case they are not in line with the requirements detailed in section 4.2 "Eligible projects" and/or fall in any of the following categories:

- proposals not considering the generation of new knowledge for the common good;
- research projects solely aimed at the development of datasets, software, new mathematical models and predictive analysis tools without appropriate consideration of the call's scientific priorities detailed in section 3 "Aims";
- proposals aimed to set up new research centres.

4.5 Submission process

Proposals must be submitted by 5:00 p.m. on **January 30th, 2024**.

5. AVAILABLE BUDGET, ELIGIBLE/NON-ELIGIBLE COSTS

The budget available for this call is 800.000 Euros.

Eligible and non-eligible costs are detailed in the previous section 4.2 "Eligible projects".

6. COMMUNICATION AND INFORMATION

Fondazione Cariplo, as a private body, is not required to comply with public procedures and has full power to decide on the allocation of the budget. The text of the call and all the mentioned

documents are available at Fondazione Cariplo website (www.fondazionecariplo.it).

For the purposes of peer review process, any personal data reported in the Project form may be transferred to non-EU Countries or organizations, in compliance with EU Regulation 2016/679. For more information, before submitting the application, please read the privacy policy on the processing of personal data and transfer of personal data to Countries or organizations outside the European Union.

7. SUMMARY

Call	Data Science for Health
Type	Deadline call
Deadline	January 30th, 2024
Available budget	€ 800.000
Aim	The call aims to support research projects that apply Data Science techniques and methods to analyze the representativeness of patients in clinical trials to develop safer and more effective treatments for all
Eligible subjects	Public or private non-profit organizations conducting scientific research
Funding costraint	Maximum funding € 200.000
Contacts	Scientific Research Area Staff contacts at: www.fondazionecariplo.it

** The data shown in the "Summary" are intended as a brief overview of the terms and conditions of the Call only. For a full description, please read carefully the text of the Call.*

8. APPENDIX: STRUCTURE AND INFORMATION ABOUT THE DATASET PROVIDED BY NOVARTIS FARMA S.p.A.

1. Introduction

With the aim of building a research tool to be shared for public benefit, Novartis Pharma S.p.A. has extracted data related to patients enrolled in a series of clinical trials coordinated by Novartis medical department and related to a variety of diseases, in complete respect for data privacy.

26 studies of both phase IIIb and phase IV were selected from the archives of Novartis medical department (19 interventional and 7 observational studies); information about approximately 15.000 patients was extracted and collected into a single dataset in an anonymized form. Selected studies share the following features:
- concluded studies;
- studies involving more than 100 patients;
- studies with *First Patient First Visit* (FPFV) later than January 1st, 2000.

2. Variables considered

Demographics

AGE: age of the patient at the time of the baseline visit. In those cases where this variable was not present, it was derived as the difference in years between the date of birth and the date corresponding to the baseline visit.

SEX: patient's sex (M, F).

RACE: patient's ethnicity (Asian, Black, Caucasian, Native American, Other, Pacific Islander, Unknown).

REGION: Italian region of the center where the study was performed and in which the patient participated. This data was derived from the province of the center with the aim of providing a geographical indication related to the patient. In compliance with privacy regulations, no data regarding patients' residence are available.

DATE: month and year of the baseline visit.

Vital signs

DBP (Dyastolic Blood Pressure): diastolic pressure (mmHg) measured during the baseline visit.

SBP (Systolic Blood Pressure): sysstolic pressure (mmHg) measured during the baseline visit.

HR (Heart Rate): number of beats per minute (bpm) measured during the baseline visit.

BMI (Body Mass Index): body mass index (Kg/m^2). In case of unavailability, this data was derived from the ratio of the patient's weight expressed in kilograms to the square of the height expressed in meters.

Laboratory tests

VAR: code/abbreviation corresponding to the laboratory variable (e.g.: "ALB" = Albumin).

LABEL: extended name of the laboratory variable.

LABVALUE: numerical value of the laboratory evaluation.

UNIT: measurement unit.

Within the database, each patient is matched with multiple records of laboratory values, according to the available laboratory data. Therefore, the number of records corresponding to a single patient depends upon the amount of laboratory evaluations that were performed at the baseline.

The overall database shows a total of 144.623 records related to 15.620 patients.

VAR1	Label	Evaluation	Unit
ALB	Albumin	Baseline	%, g/L, g/dL, mg/dL
ALP	ALP	Baseline	U/L
ALT	ALT	Baseline	U/L
AST	AST	Baseline	U/L
BAS	Basophils	Baseline	%, 10^3/mL, 10^3/µL, 10^9/L
CA	Calcium	Baseline	mEq/L, mg/dL, mmol/L
CHOL	Total cholesterol	Baseline	mg/dL, mmol/L
CL	Chlorine	Baseline	mEq/L, mmol/L
CREA	Serum Creatinine	Baseline	mg/dL, µmol/L
DBIL	Direct bilirubin	Baseline	mg/dL
EOS	Eosinophils	Baseline	%, 10^3/mL, 10^3/µL, 10^9/L
GLY	Glycemia	Baseline	mg/dL, mmol/L
HBA1C	Glycated hemoglobin	Baseline	%
HCT	Hematocrit	Baseline	%
HDL	HDL	Baseline	mg/dL, mmol/L
HGB	Hemoglobin	Baseline	g/dL, g/L
K	Potassium	Baseline	mEq/L, mmol/L
LDL	LDL	Baseline	mg/dL, mmol/L
LYM	Lymphocytes	Baseline	%, 10^3/mL, 10^3/µL, 10^9/L
MON	Monocytes	Baseline	%, 10^3/mL, 10^3/µL, 10^9/L
N	Nitrogen	Baseline	mg/dL
NEU	Neutrophils	Baseline	%, 10^3/mL, 10^3/µL, 10^9/L
OTH	Other	Baseline	%, 10^9/L
PLAT	Platelets	Baseline	10^3/µL, 10^9/L
RBC	Erythrocytes	Baseline	10^12/L, 10^3/µL, 10^6/L, 10^6/µL
SOD	Sodium	Baseline	mmol/L, mEq/L
TBIL	Total Bilirubin	Baseline	mg/dL, µmol/L
TPROT	Total protein	Baseline	g/L, g/dL
TRI	Triglycerides	Baseline	mg/dL, mmol/L
UACID	Urate	Baseline	mg/dL, mmol/L, µmol/L
UREA	Urea	Baseline	mg/dL, g/L
WBC	Leukocytes	Baseline	10^3/µL, 10^9/L

Laboratory tests reference values

Reference values may vary depending on age, sex and the specific instrumentation in use in each center. Therefore, reference values may be different both between different centers and within the same center.

LORNG: lower limit of the reference range. A numerical value of the laboratory test smaller than the lower limit may be considered as altered.

HIRNG: upper limit of the reference range. A numerical value of the laboratory test bigger than the upper limit may be considered as altered.

Added variables

SUBJID: patient ID. Random number assigned to each patient, from 1 to 16.557. Please note that the total number of patients is 15.620; therefore, patient IDs are not consecutive.
PATCOD: pathology ID of the study from which patients' data were extracted.

3. Database overview

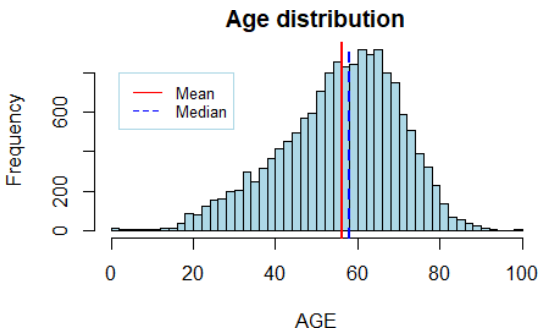
The database contains 144.623 records related to 15.620 patients, is organized in 17 columns and is displayed in "long" format (multiple records for each patient) for the variable "VAR".

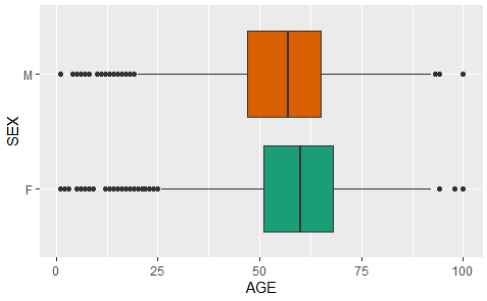
Patients

Sex: 55% male and 45% female

Age: overall, mean age is around 56 years and. Males' mean age is 55,8 years, while female's one is 56,7.

Ethnicity: almost entirely Caucasian subjects (Asian, Black, Native American subjects and Pacific islanders are lesser than 1%).





Studies' participants were enrolled mainly in centers located in Campania, Lombardy, and Sicily regions (1852, 1729 and 1449, respectively). The map below represents the distribution of patients according to the region of the centers.

Studies

The table below shows the distribution of patients according to the relevant pathology.

PATCOD	Number of patients	PATCOD	Number of patients
CNV	207	MetS	268
DIA_2	8918	PARTIAL SEIZURE	284
HBP	579	PSO_00	653
HEA_TRANS	385	PSO_01	431
KID_TRANS	2063	PSO_02	258
LIV_TRANS	186	SD	176
MS	1212		

For each pathology, studies' inclusion and exclusion criteria are available.

Limits

The dataset presents some limitations related to the variety of the characteristics of the considered studies:

- the inclusion of observational studies increased the number of observations but generated a large number of missing values;
- laboratory evaluations are derived from different Italian centers and, therefore, are reported with different measurement units.