Machine learning applied to electronic health record data: opportunities and challenges

Riccardo Bellazzi

Dipartimento di Ingegneria Industriale e dell'Informazione, Università di Pavia
IRCCS ICS Maugeri, Pavia

riccardo.bellazzi@unipv.it

UNIVERSITÀ DI PAVIA

A little disillusion

- Pavia ≠ Padua (Padova)

However, a University since 1361

UNIVERSITÀ DI PAVIA

And … we also have an historical anatomical theatre

Hospitals and engineering

1 mile

UNIVERSITÀ DI PAVIA

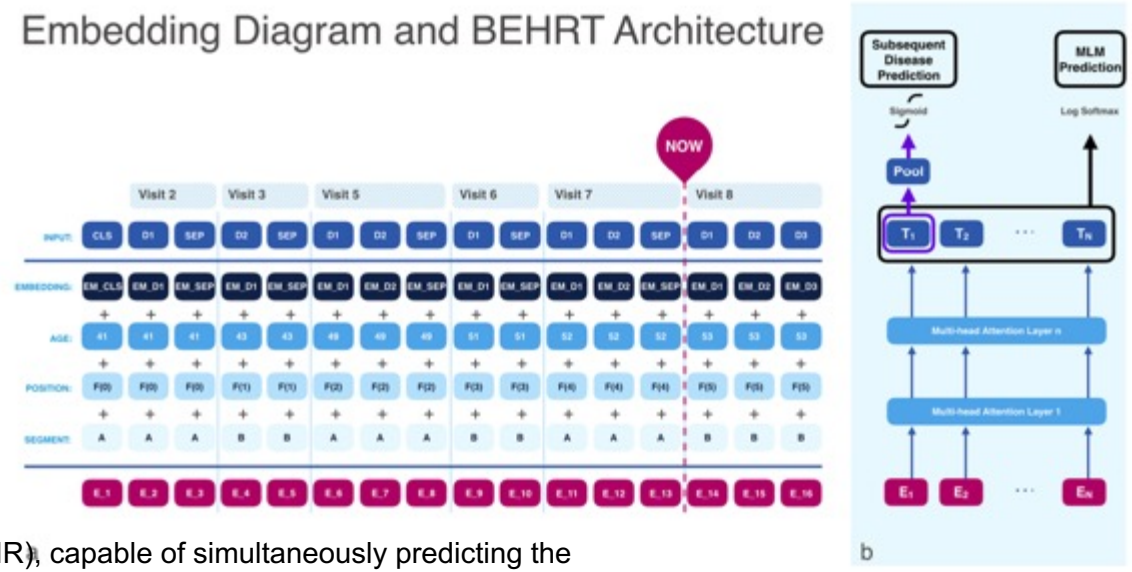
SCIENTIFIC REPORTS
nature research
Check for updates

OPEN

BEHRT: Transformer for Electronic Health Records

Yikuan Li[1,2], Shishir Rao[1,2], José Roberto Ayala Solares[1], Abdelaali Hassaine[1], Rema Ramakrishnan[1], Dexter Canoy[1], Yajie Zhu[1], Kazem Rahimi[1] & Gholamreza Salimi-Khorshidi[1]

Sci Rep. 2020; 10: 7155.
Published online 2020 Apr 28. doi: 10.1038/s41598-020-62922-y



BEHRT: A deep neural sequence transduction model for electronic health records (EHR), capable of simultaneously predicting the likelihood of 301 conditions in one's future visits.

Patterns

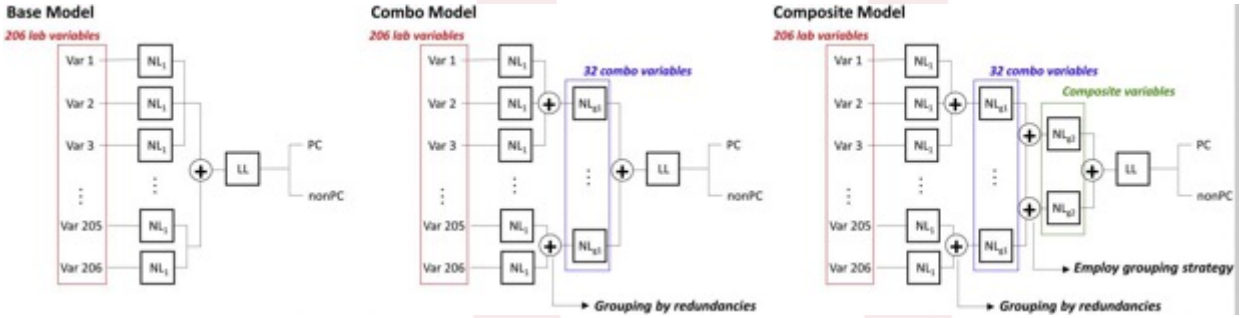
CellPress
OPEN ACCESS

Article

Structured deep embedding model to generate composite clinical indices from electronic health records for early detection of pancreatic cancer

Jiheum Park,[1,9] Michael G. Artin,[2] Kate E. Lee,[3] Benjamin L. May,[4] Michael Park,[5,6] Chin Hur,[1,8,*] and Nicholas P. Tatonetti[7,8]
[1]Department of Medicine, Columbia University Irving Medical Center, New York, NY 10032, USA
[2]Hospital of the University of Pennsylvania, Philadelphia, PA 19104, USA
[3]Duke University Medical Center, Durham, NC 27710, USA
[4]Herbert Irving Comprehensive Cancer Center, Columbia University Irving Medical Center, New York, NY 10032, USA
[5]Applied Info Partners, Inc, Worlds Fair Drive, Somerset, NJ 08873, USA
[6]X-Mechanics, Cresskill, NJ 07626, USA
[7]Department of Biomedical Informatics, Columbia U
[8]Senior author
[9]Lead contact
*Correspondence: ch447@cumc.columbia.edu
https://doi.org/10.1016/j.patter.2022.100636

Park et al., 2023, Patterns 4, 100636
January 13, 2023
https://doi.org/10.1016/j.patter.2022.100636



In this study, we developed our model with laboratory measurement data, in contrast to most models based on EHR data that rely primarily on ICD codes.

UNIVERSITÀ DI PAVIA

Towards trustworthy systems (2019)

Seven requirements for implementation of AI trustworthy solutions:

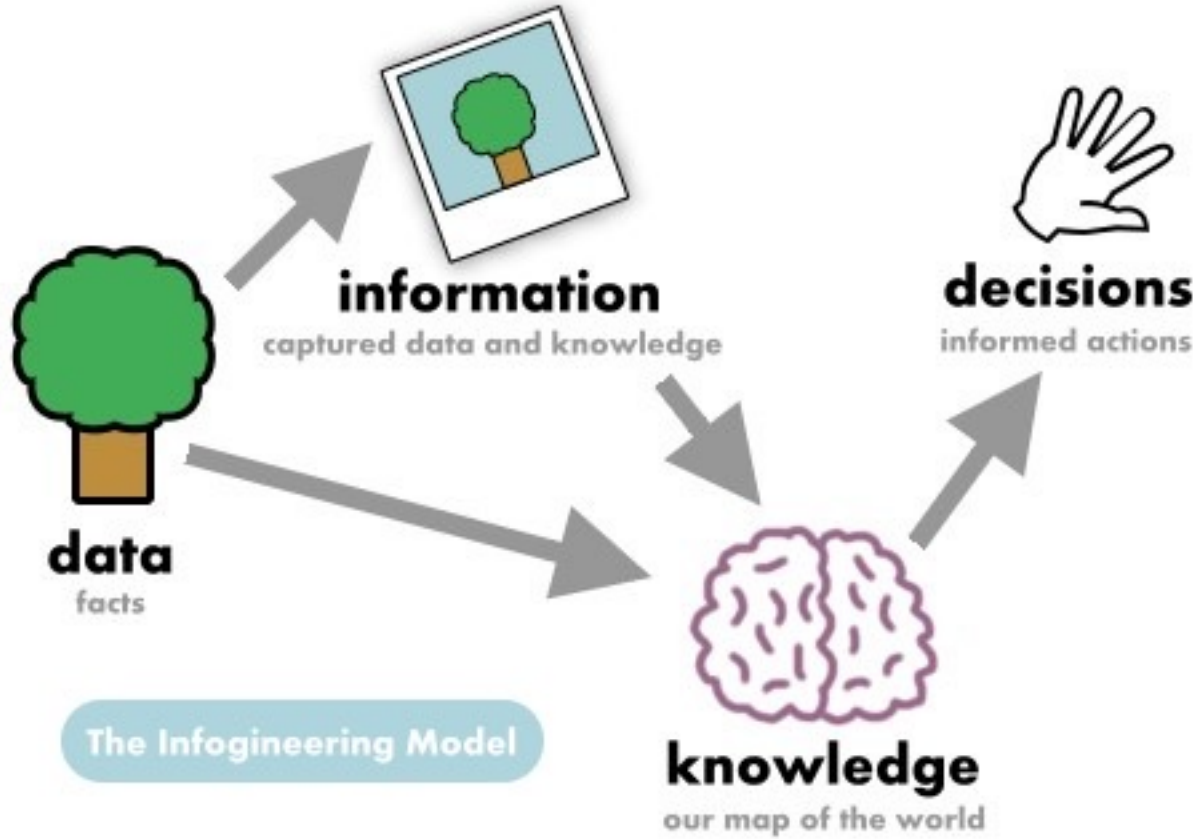
- human agency and oversight
- Transparency
- technical robustness and safety
- privacy and data governance
- diversity non-discrimination and fairness
- societal and environmental well-being
- accountability

What makes trustworthy an AI system based on EHR data?

UNIVERSITÀ DI PAVIA

Is AI on EHR data the «pipe piper»?

Knowledge is in the data - Data are key




INTERNATIONAL JOURNAL OF MEDICAL INFORMATICS 82 (2013) 1–9

ELSEVIER

journal homepage: www.ijmijournal.com

Trustworthy reuse of health data: A transnational perspective

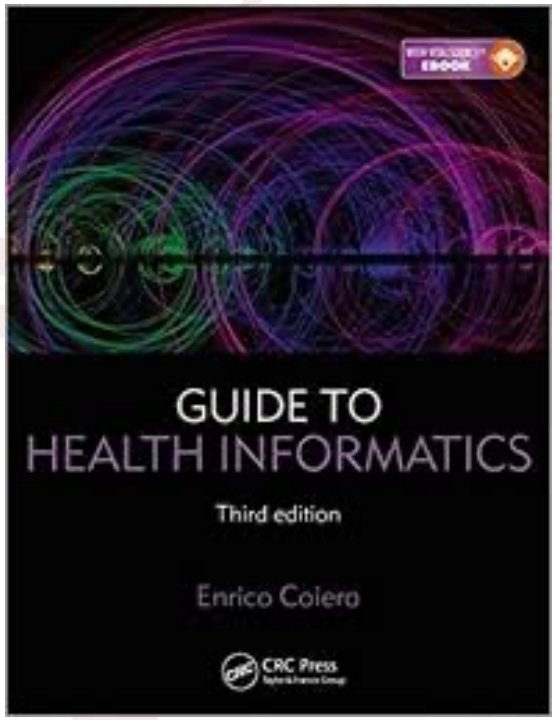
A. Geissbuhler[a,b,*], C. Safran[c], I. Buchan[d], R. Bellazzi[a,e], S. Labkoff[f], K. Eilenberg[g], A. Leese[h], C. Richardson[i], J. Mantas[a,j], P. Murray[a], G. De Moor[k]

Needs:

- Data governance
- Data curation and Stewardship

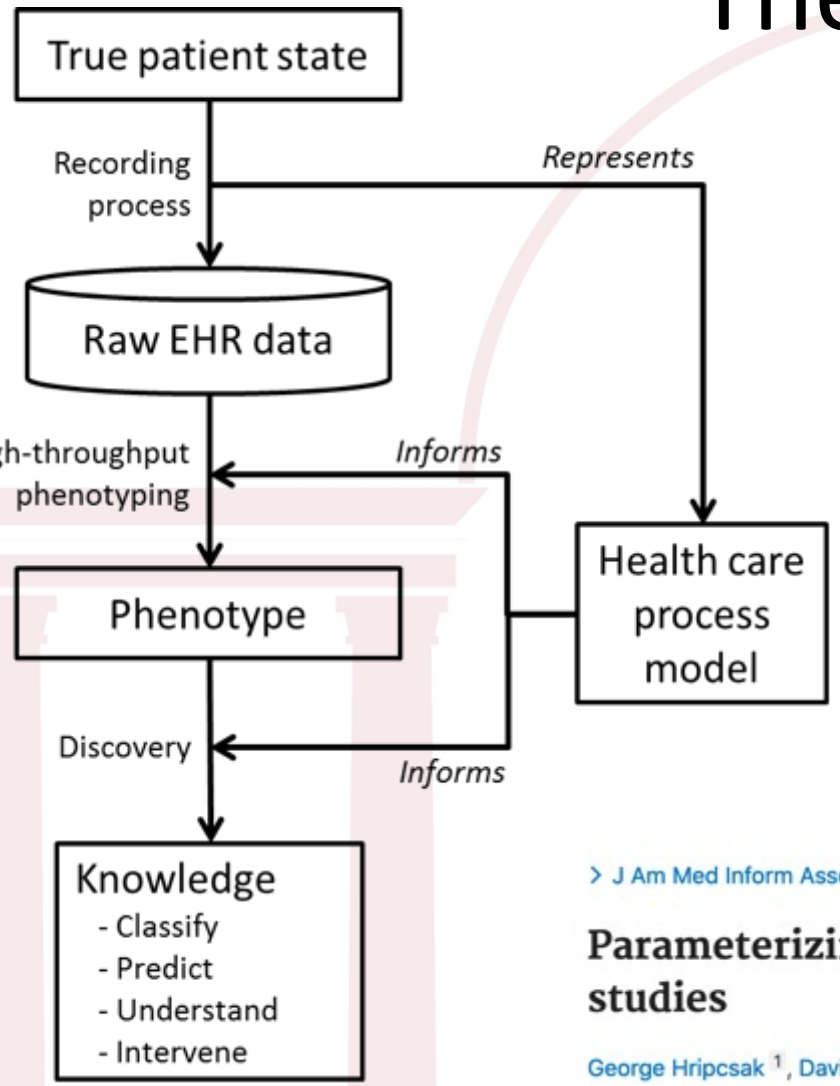
http://www.infogineering.net/data-information-knowledge.htm

Understanding the context



Health information systems are purposive

The process of care



J Am Med Inform Assoc. 2015 Jul;22(4):794-804. doi: 10.1093/jamia/ocu051. Epub 2015 Feb 26.

Parameterizing time in electronic health record studies

George Hripcsak [1], David J Albers [2], Adler Perotte [2]

A worldwide challenge – COVID pandemics

Consortium for Clinical Characterization of COVID-19 by EHR

https://covidclinical.net/

March 2020 Consortium Formed

Gather key questions and identify data-driven approaches for studying the COVID-19 pandemic, leveraging EHR systems and the i2b2 community.

April 2020. First preprint publication.

Today. 37 Members. NIH funding request.
More than 10 journal and preprint publications

342 hospitals

8 countries

37,000 patients

UNIVERSITÀ DI PAVIA

4CE - Italy

Istituti Clinici Scientifici Maugeri

Ospedale di Bergamo

Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico

Sistema Socio Sanitario
Regione Lombardia
ASST Pavia

BIOMERIS

AZIENDA OSPEDALIERO UNIVERSITARIA MATER DOMINI

UNIVERSITÀ DI PAVIA

Two Phases

A federated model based on locally-run analyses enables 4CE to "stay close to the data"

Nation Diversity, Global Perspectives
(Phase 1.x)

- **342 hospitals, 8 countries, 37,000 patients admitted for COVID-19**
- **Different hospital perspectives, regional and country variation**
- **SQL queries run on i2b2, OMOP, and others; leverages ACT ontology**
- **Low regulatory barriers to participation**

Federated Model

- Analyses run **locally** at sites, only share aggregate counts and statistics
- **Local** data experts and clinicians refine questions, know coding practices, perform chart review
- Data quality problems can be fixed in **local** databases

Deep Analysis, Chart Review
(Phase 2.x)

- Project-specific subsets of sites (not all sites needed)
- Deep dives into sites' data with chart review to validate data and methods
- ML models and complex analyses using R on patient-level data
- Run on Docker image to create a standardized compute environment

UNIVERSITÀ DI PAVIA

(thanks to G. Weber)

Two Phases

A federated model based on locally-run analyses enables 4CE to "stay close to the data"

Nation Diversity, Global Perspectives
(Phase 1.x)

- 342 hospitals, 8 countries, 37,000 patients admitted for COVID-19
- Different hospital perspectives, regional and country variation
- SQL queries run on i2b2, OMOP, and others; leverages ACT ontology
- Low regulatory barriers to participation

Federated Model

- **Analyses run locally at sites, only share aggregate counts and statistics**
- **Local data experts and clinicians refine questions, know coding practices, perform chart review**
- **Data quality problems can be fixed in local databases**

Deep Analysis, Chart Review
(Phase 2.x)

- Project-specific subsets of sites (not all sites needed)
- Deep dives into sites' data with chart review to validate data and methods
- ML models and complex analyses using R on patient-level data
- Run on Docker image to create a standardized compute environment

UNIVERSITÀ DI PAVIA

(thanks to G. Weber)

Two Phases

A federated model based on locally-run analyses enables 4CE to "stay close to the data"

Nation Diversity, Global Perspectives
(Phase 1.x)

- 342 hospitals, 8 countries, 37,000 patients admitted for COVID-19
- Different hospital perspectives, regional and country variation
- SQL queries run on i2b2, OMOP, and others; leverages ACT ontology
- Low regulatory barriers to participation

Federated Model

- Analyses run **locally** at sites, only share aggregate counts and statistics
- **Local** data experts and clinicians refine questions, know coding practices, perform chart review
- Data quality problems can be fixed in **local** databases

Deep Analysis, Chart Review
(Phase 2.x)

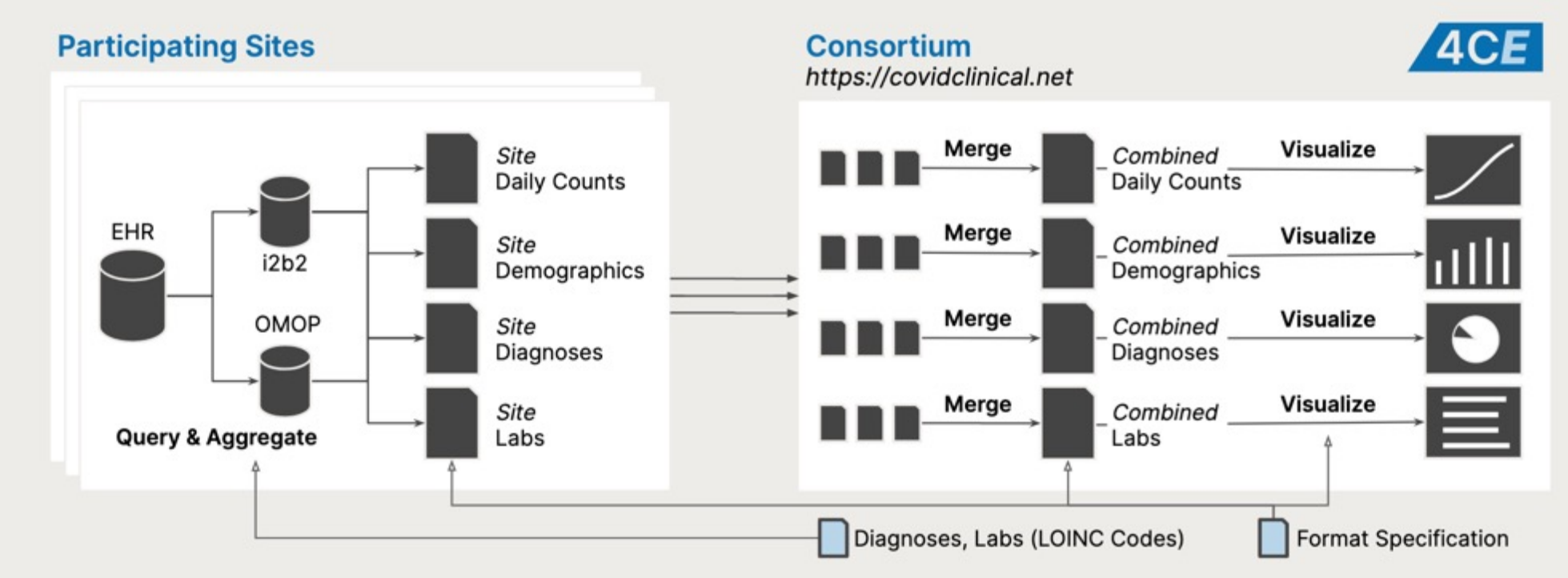
- **Project-specific subsets of sites (not all sites needed)**
- **Deep dives into sites' data with chart review to validate data and methods**
- **ML models and complex analyses using R on patient-level data**
- **Run on Docker image to create a standardized compute environment**

UNIVERSITÀ DI PAVIA

(thanks to G. Weber)

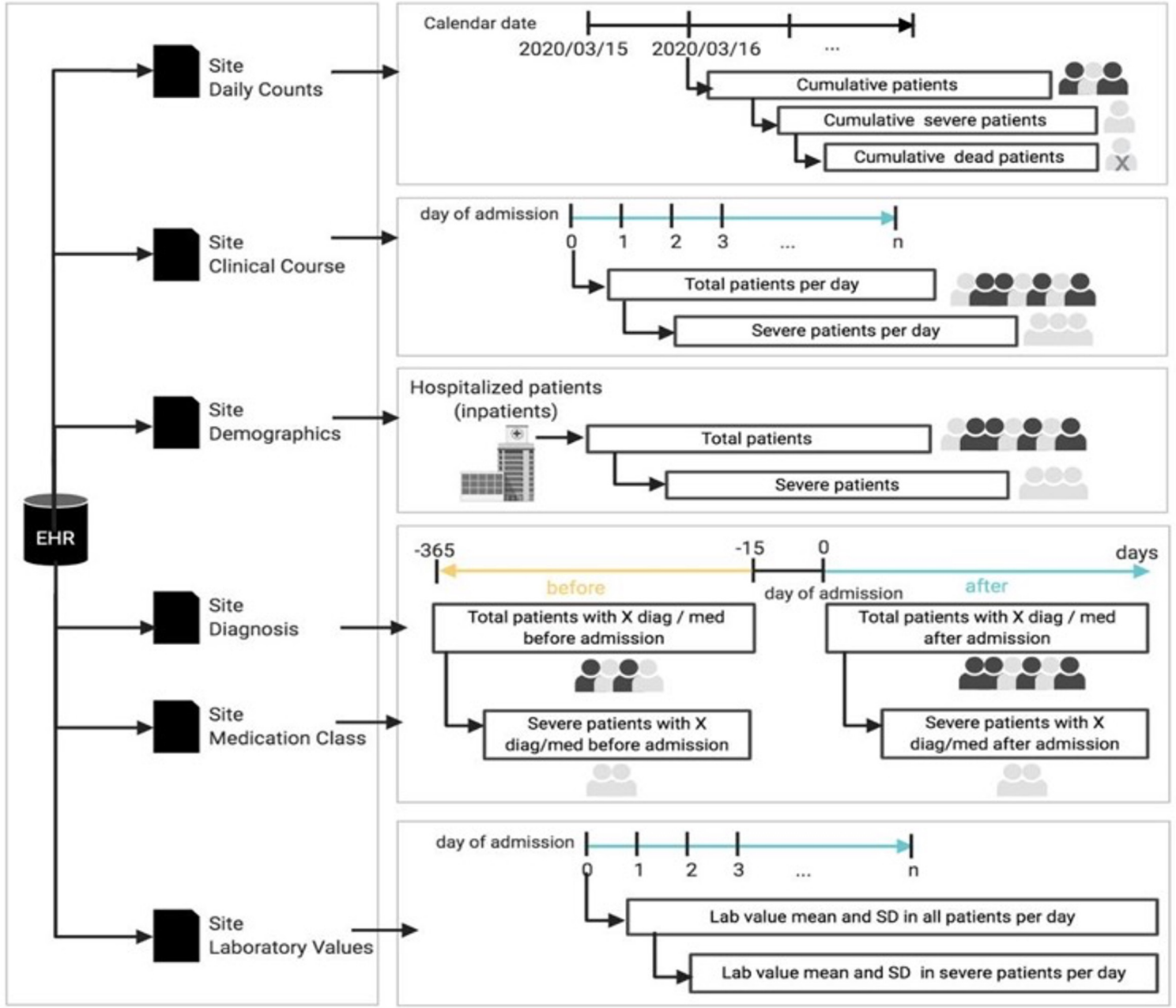
Phase 1. Data collection and analysis

Upload aggregate data, quality checks



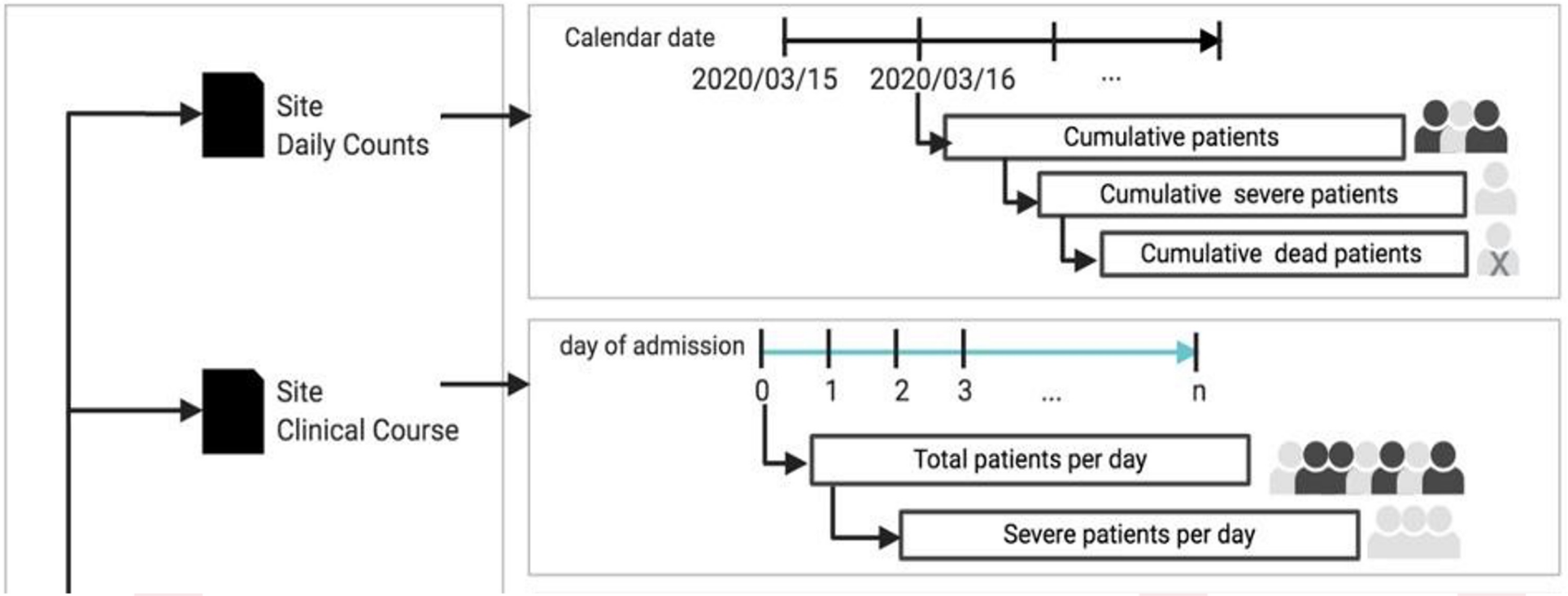
UNIVERSITÀ DI PAVIA

Brat, G.A., Weber, G.M., Gehlenborg, N. et al. International electronic health record-derived COVID-19 clinical course profiles: the 4CE consortium. npj Digit. Med. 3, 109 (2020). https://doi.org/10.1038/s41746-020-00308-0



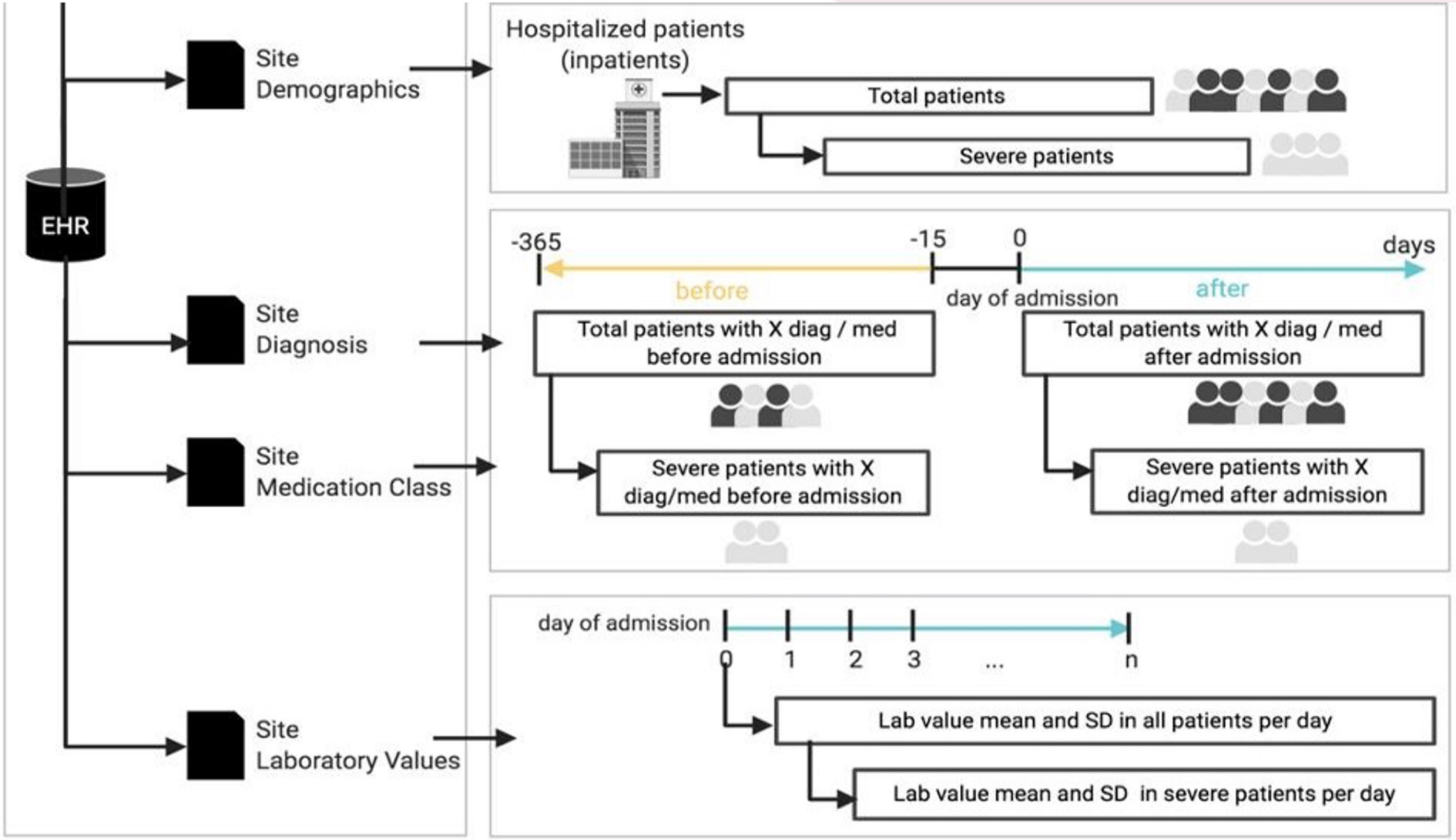
UNIVERSITÀ DI PAVIA

Brat, G.A., Weber, G.M., Gehlenborg, N. et al. International electronic health record-derived COVID-19 clinical course profiles: the 4CE consortium. npj Digit. Med. 3, 109 (2020). https://doi.org/10.1038/s41746-020-00308-0



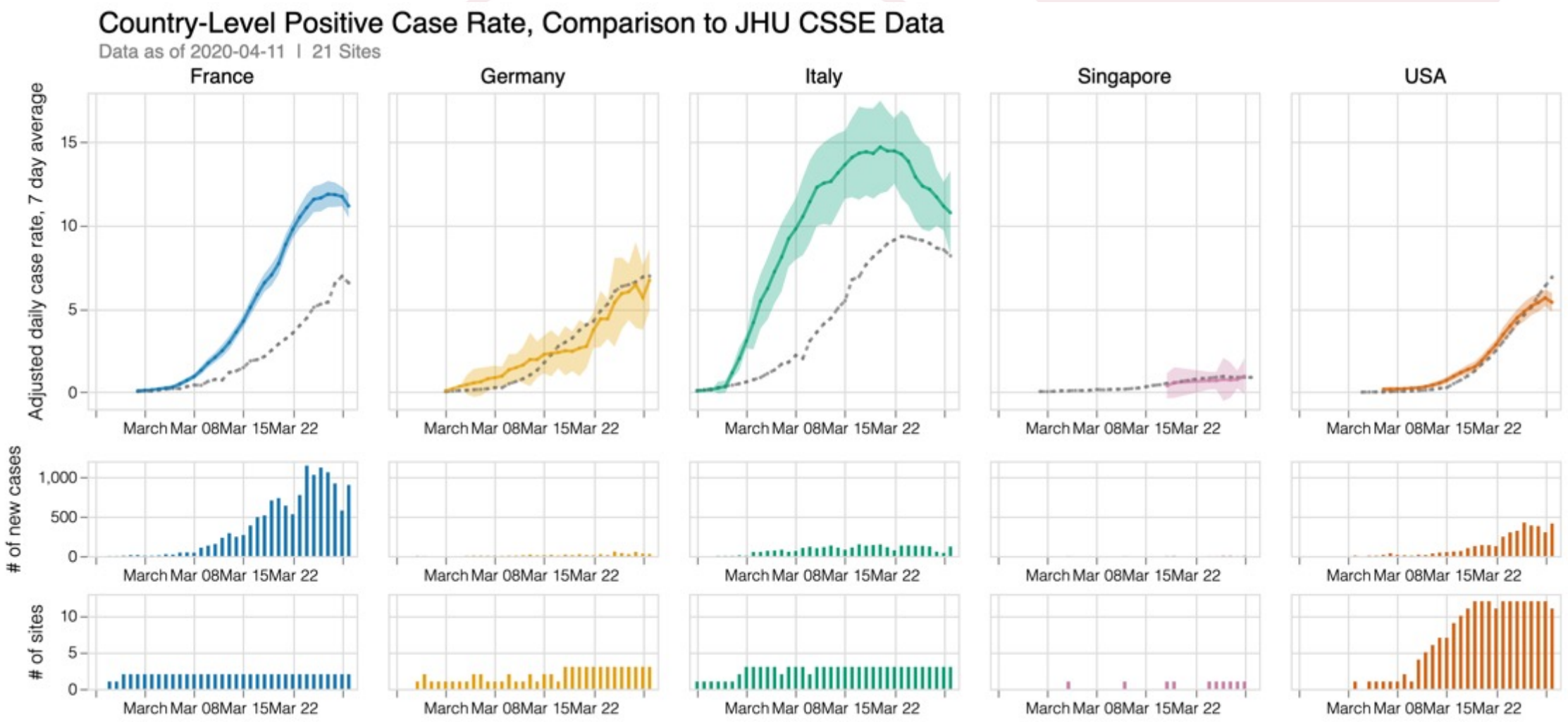
UNIVERSITÀ DI PAVIA

Brat, G.A., Weber, G.M., Gehlenborg, N. et al. International electronic health record-derived COVID-19 clinical course profiles: the 4CE consortium. npj Digit. Med. 3, 109 (2020). https://doi.org/10.1038/s41746-020-00308-0



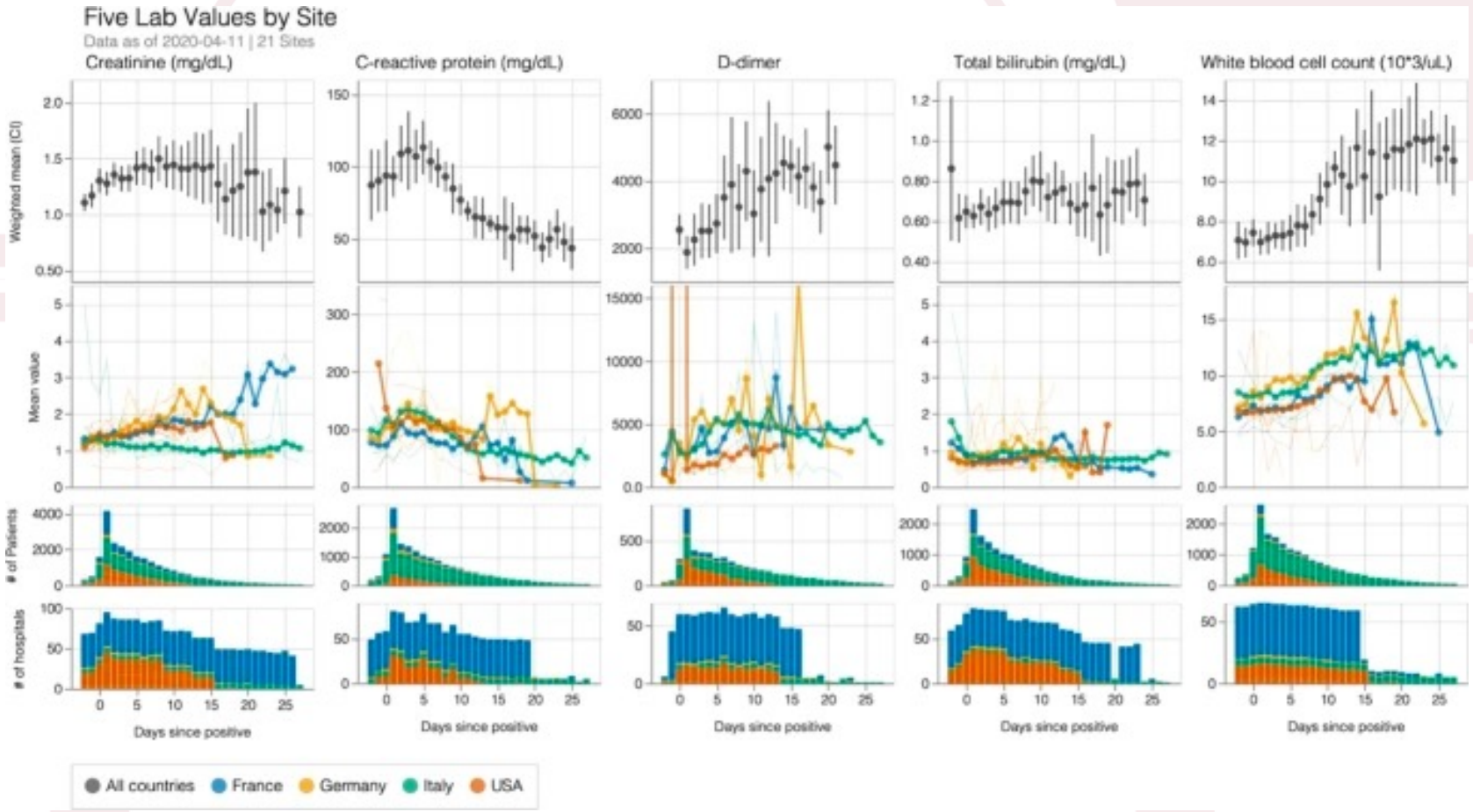
Brat, G.A., Weber, G.M., Gehlenborg, N. et al. International electronic health record-derived COVID-19 clinical course profiles: the 4CE consortium. npj Digit. Med. 3, 109 (2020). https://doi.org/10.1038/s41746-020-00308-0

Phase 1 Results, 1° wave (March 2020)



UNIVERSITÀ DI PAVIA

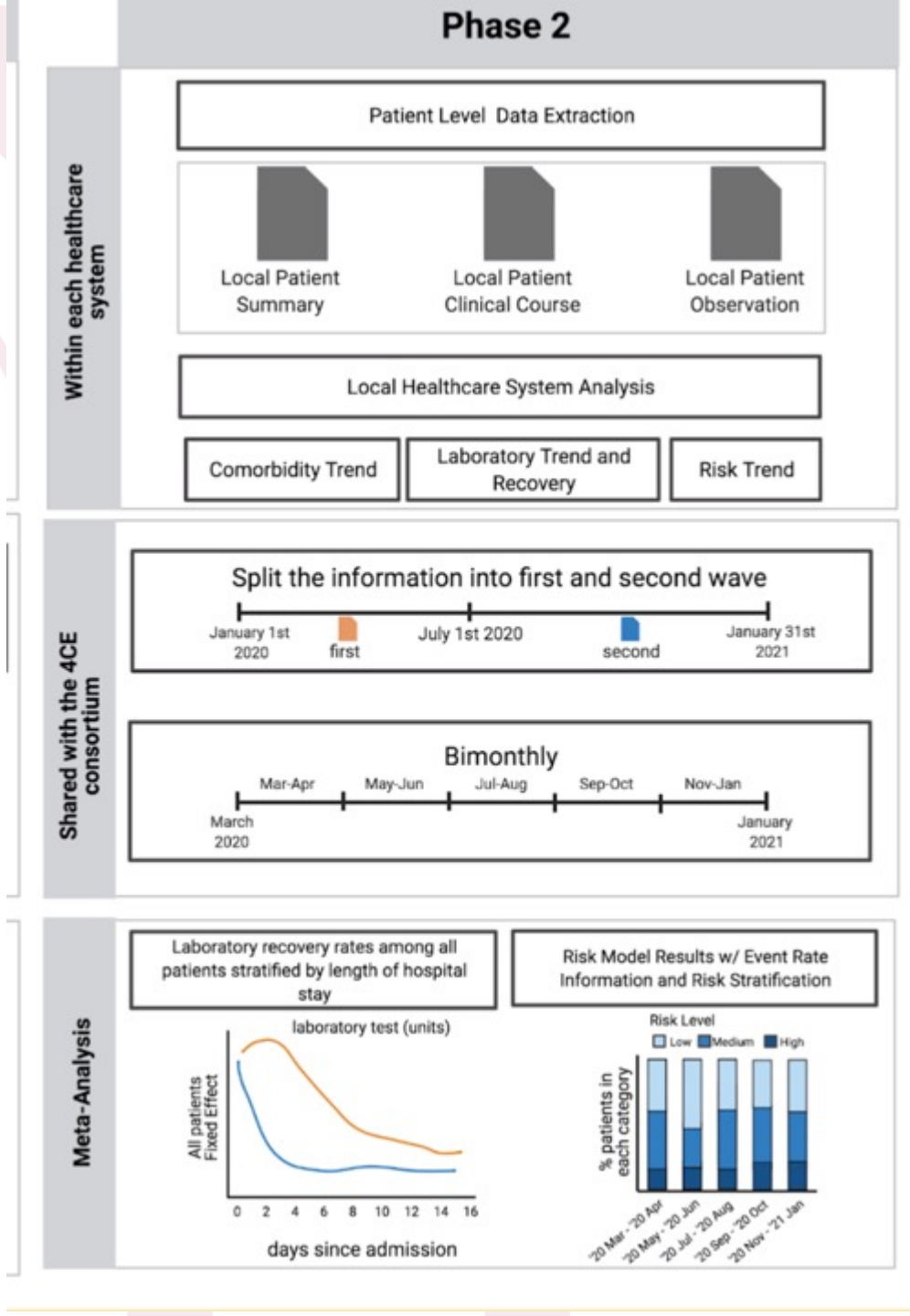
Laboratory tests representative of renal function (creatinine), systemic inflammation (C-reactive protein), coagulopathy (D-dimer), liver function (total bilirubin), and immune response (white blood cell count) visualized relative to date of diagnosis of COVID-19.



Phase 2 of the Consortium

Patient Level Analyses

R scripts run on a Docker image at each hospital to provide a standardized local computing environment (still only share aggregate results externally)



UNIVERSITÀ DI PAVIA

From basic score validation to federated learning

UNIVERSITÀ DI PAVIA

Severity Score

Validation of an Internationally Derived Patient Severity Phenotype to Support COVID-19 Analytics from Electronic Health Record Data

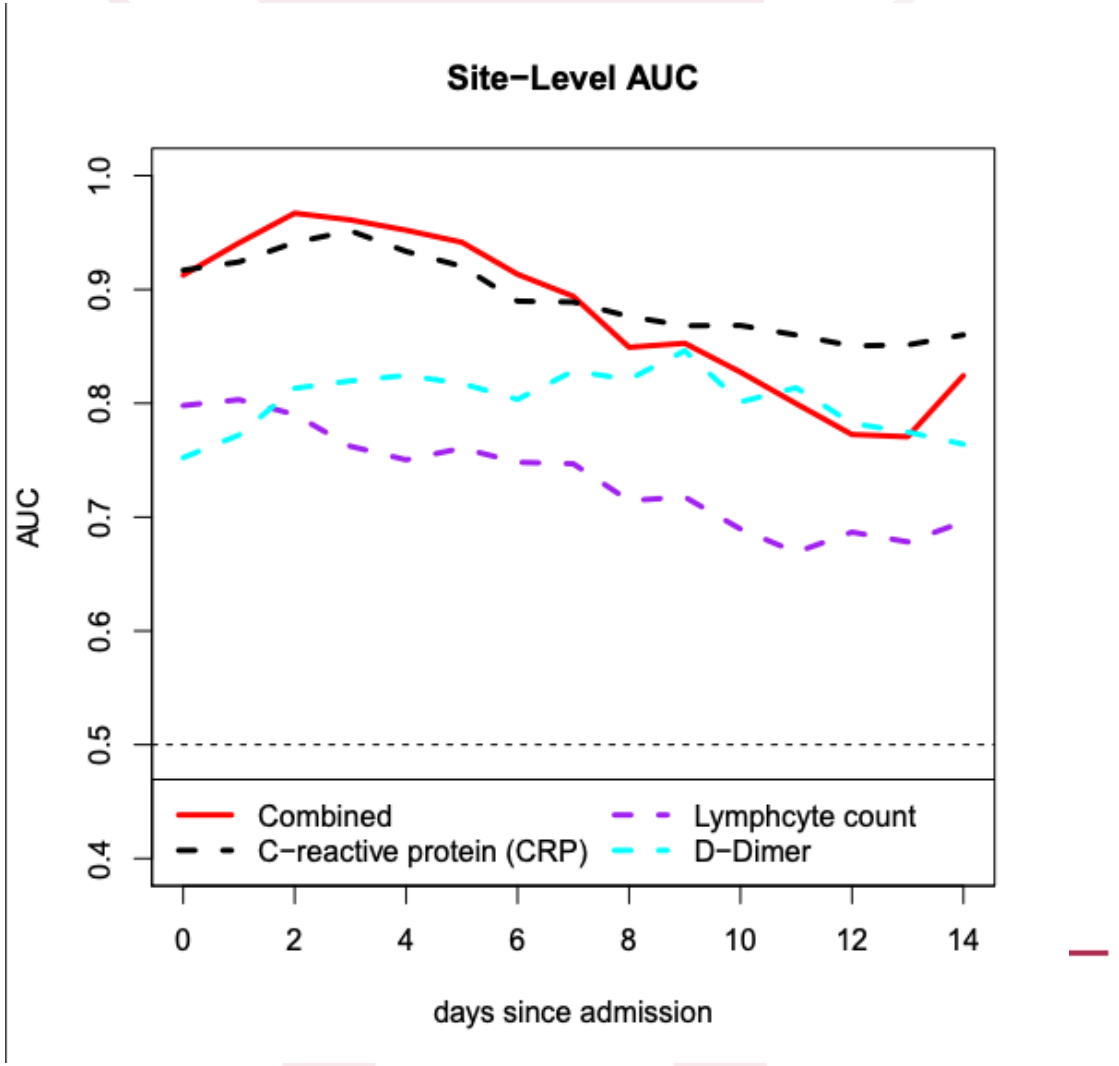
J. Klann et al, 2021, JAMIA

A 4CE severity phenotype that is both clinically reasonable and possible to identify across our diverse sites.

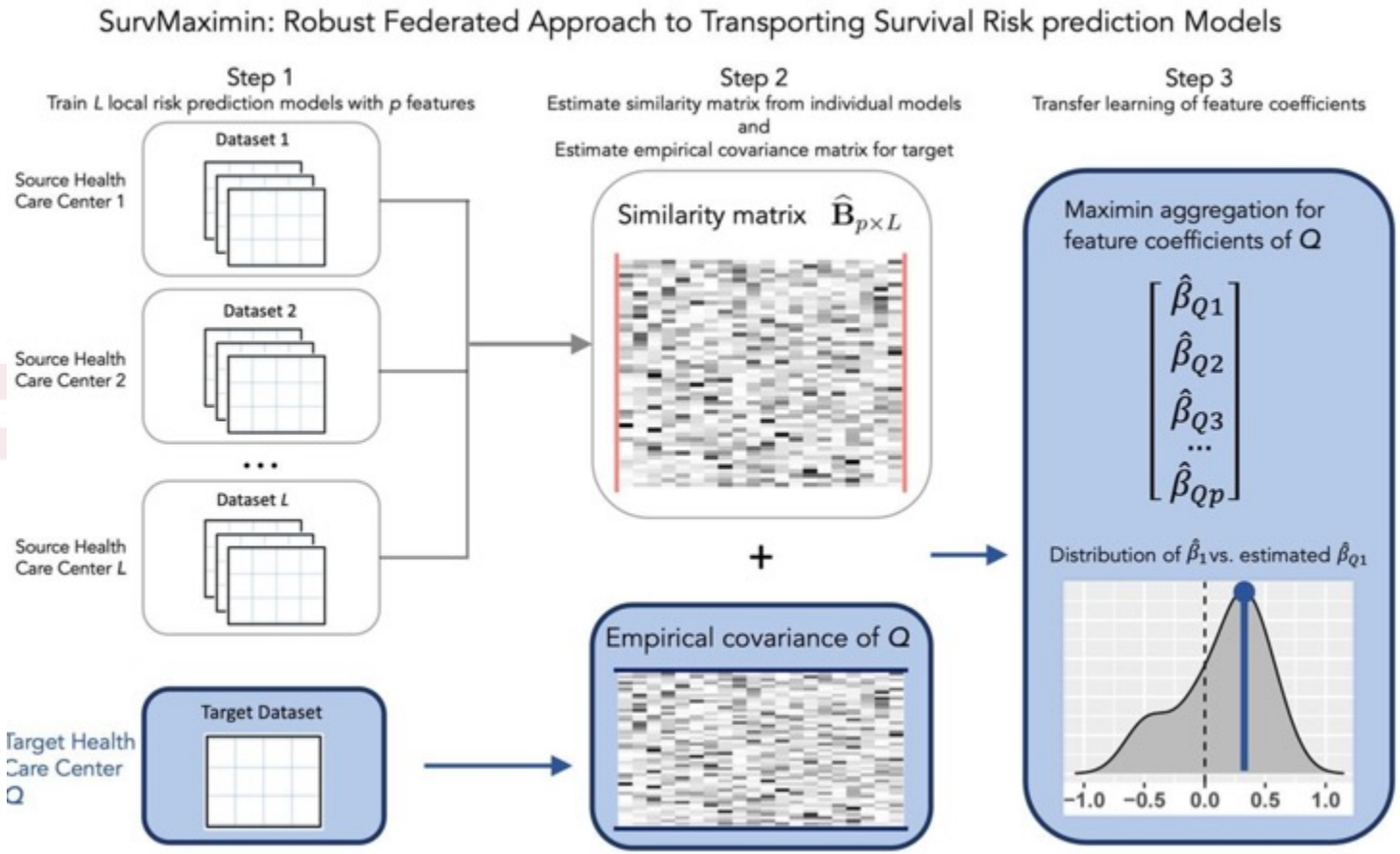
Limit severity to the EHR data classes that 4CE is collecting:

demographics, diagnoses, medications, labs, and ICD procedure codes.

We did not use outcomes (e.g., ICU admission), symptoms (e.g., wheezing), or vital signs (e.g., respiratory rate), as these are not widely or reliably available in EHRs.



UNIVERSITÀ DI PAVIA



ELSEVIER

Journal of Biomedical Informatics

Volume 134, October 2022, 104176

Original Research

SurvMaximin: Robust federated approach to transporting survival risk prediction models

Xuan Wang [a], Harrison G. Zhang [b], Xin Xiong [a], Chuan Hong [b], Griffin M. Weber [b], Gabriel A. Brat [b], Clara-Lea Bonzel [b], Yuan Luo [c], Rui Duan [d], Nathan P. Palmer [b], Meghan R. Hutch [c], Alba Gutiérrez-Sacristán [b], Riccardo Bellazzi [e], Luca Chiovato [f], Kelly Cho [g h], Arianna Dagliati [e], Hossein Estiri [i], Noelia García-Barrio [j], Romain Griffier [k l], David A. Hanauer [m]...Tianxi Cai [b 1]

UNIVERSITÀ DI PAVIA

Working Groups and projects

- Projects to refine and validate methods
  - COVID-19 disease severity algorithm
  - Longitudinal analyses: differences between pandemic waves
- Projects looking at understudied or underrepresented populations
  - Pediatrics
  - Race and ethnicity
- Projects on disease-specific diagnosis, risk factors, management and outcomes
  - Neurological diseases
  - Acute kidney injury
  - Thrombotic events

Two international initiatives made this possible

- **OHDSI** (Observational Health Data Sciences and Informatics)
  - Based on OMOP database
  - +2500 active collaborators worldwide
  - +400 healthcare organizations
  - +74 countries
  - +800 millions unique patient records
  - approx. 11% world's population
  - Billions of patients' facts
- **I2B2/TRANSMART**
  - Informatics for Integrating Biology & the Bedside, or i2b2 is an NIH-funded enterprise clinical research platform, which contains a database model, application layer, and core APIs.

Viewpoint

What Every Reader Should Know About Studies Using Electronic Health Record Data but May Be Afraid to Ask

(J Med Internet Res 2021;23(3):e22219) doi: 10.2196/22219

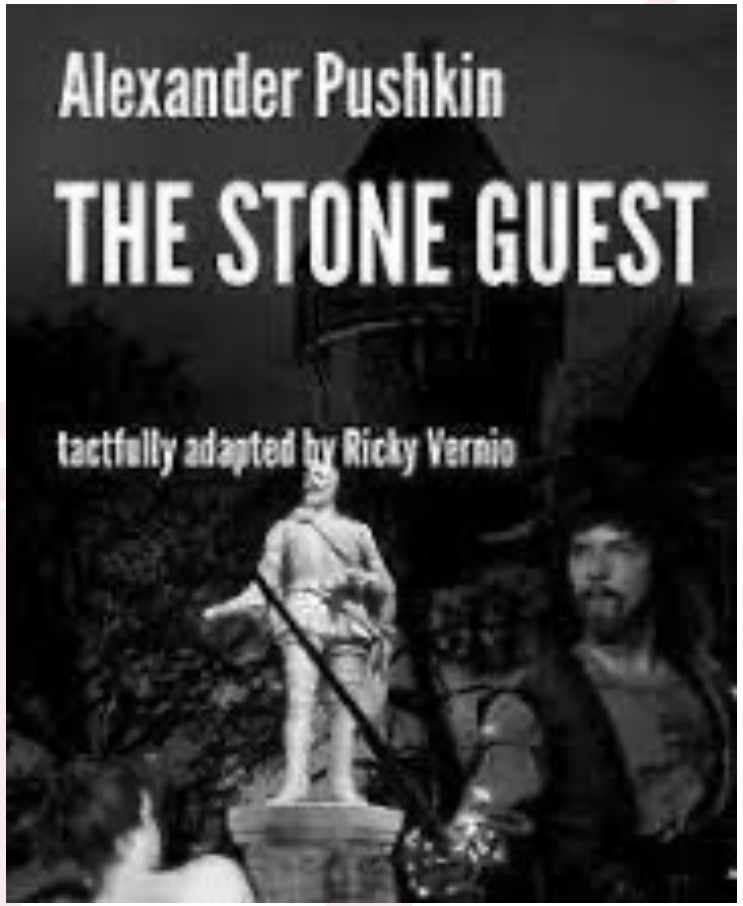
1. *How complete are the data?*
2. *How were the data collected and handled?*
3. *What were the specific data types?*
4. *Did the analysis account for EHR variability?*
5. *Are the data and analytic code transparent?*
6. *Was the study appropriately multidisciplinary?*

Data Completeness, Data harmonization and handling, Data types

- Data Types, Coding System and Data Transformation
- Deidentification Strategy
- Sparse Data and Management of Missing Values
- Data Scattered in Different Sources and Integration Strategy (including NLP)
- Data inferred and Computational Phenotyping
- Time, Time-stamps, Granularity and Validity time.
- Partial view of the patient's History

Robustness, Transparency and multidisciplinary approach

- Robustness of the analysis against EHR variability
  - Variability due to population
  - Variability due to the healthcare processes
- Code should be made available together with synthetic data / fully anonymized data
  - Explicit variable transformation strategy used for learning
  - Full preprocessing pipeline available
- Multidisciplinary approach
  - Need of a multidisciplinary view since model construction



EU Regulation n. 2016/679 or
General Data Protection Regulation
In Italy DL 101/2018

Towards trustworthy systems (2019)

Seven requirements for implementation of AI trustworthy solutions:

- human agency and oversight
- Transparency
- technical robustness and safety
- privacy and data governance
- diversity non-discrimination and fairness
- societal and environmental well-being
- accountability

Artificial Intelligence In Medicine 133 (2022) 102423

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

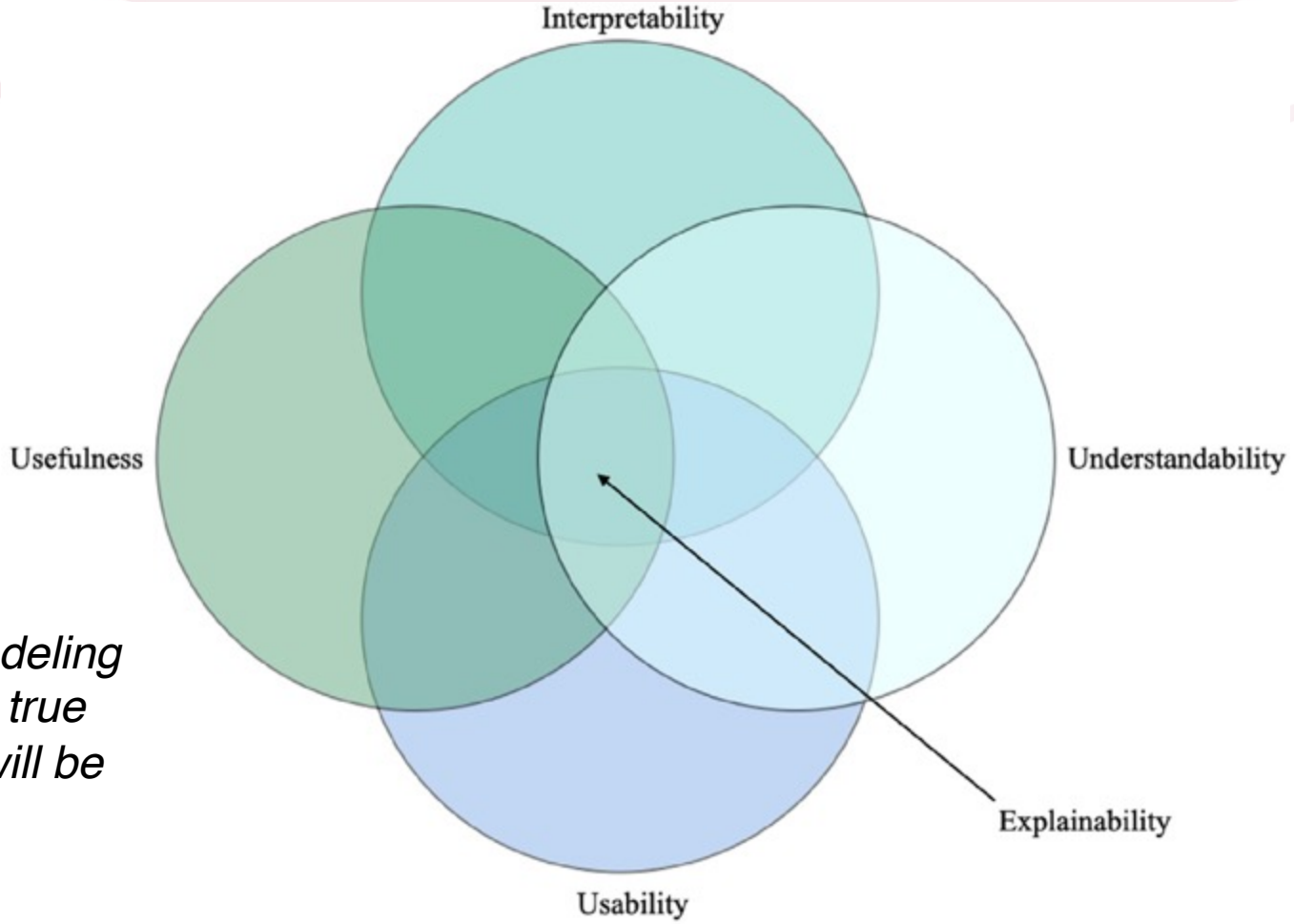
journal homepage: www.elsevier.com/locate/artmed

Research paper

A manifesto on explainability for artificial intelligence in medicine

Carlo Combi [a,*], Beatrice Amico [a], Riccardo Bellazzi [b], Andreas Holzinger [c], Jason H. Moore [d], Marinka Zitnik [e], John H. Holmes [f]

[a] University of Verona, Verona, Italy
[b] University of Pavia, Pavia, Italy
[c] Medical University Graz, Graz, Austria
[d] Cedars-Sinai Medical Center, West Hollywood, CA, USA
[e] Harvard Medical School and Broad Institute of MIT & Harvard, MA, USA
[f] University of Pennsylvania Perelman School of Medicine Philadelphia, PA, USA



Proposition: XAI-based systems need to start from modeling the biomedical and clinical domain in order to obtain a true understanding of the context in which these systems will be used.

Proposition: Explanations are not always required in order for an AI model to be useful. Functional specifications obtained from deep analysis of the problem domain and users should determine when explainability and interpretability are required.

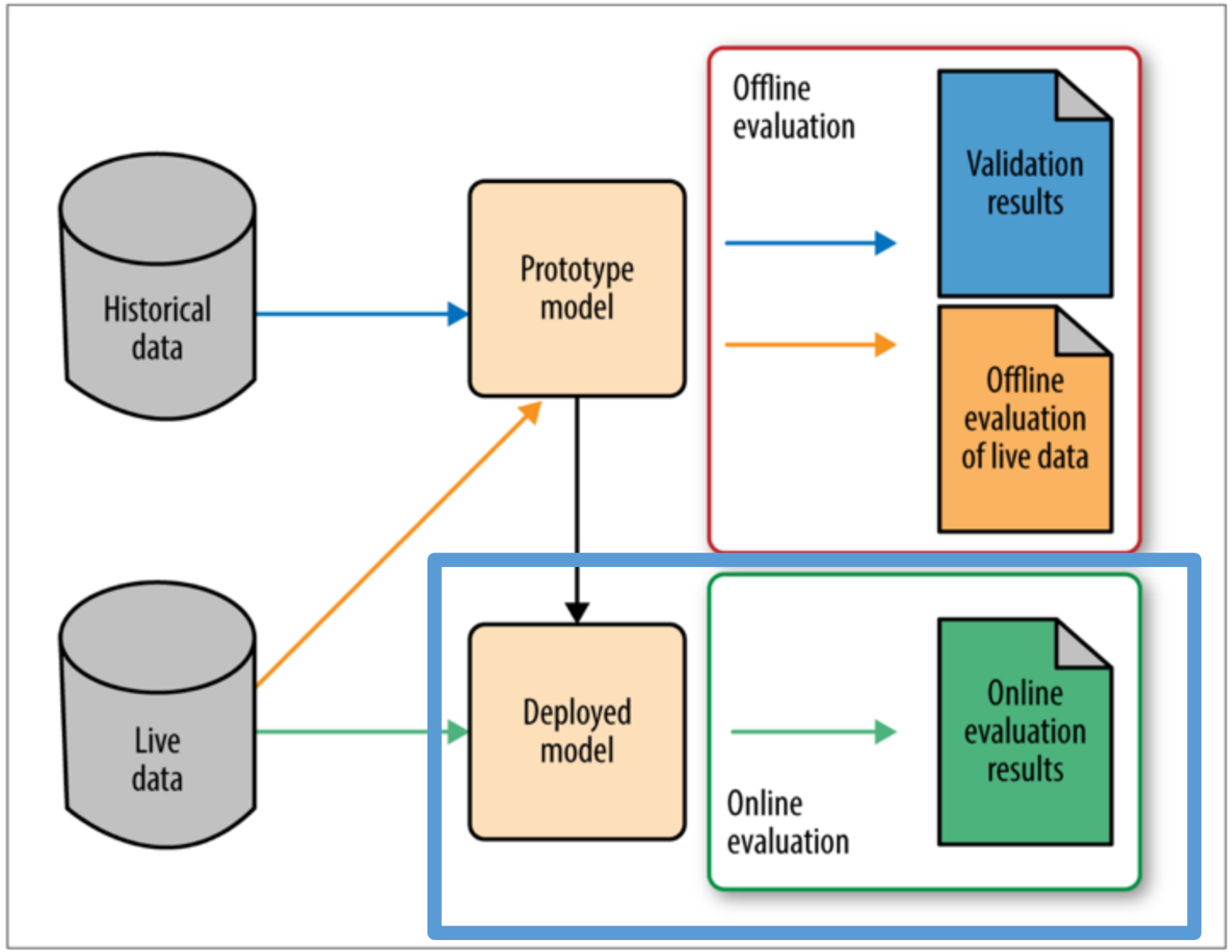


Figure 1-1. Machine learning model development and evaluation workflow

Reliability: a necessary property of Trustworthy AI

- Reliability engineering is an engineering discipline for applying scientific know-how to a component, product, plant, or process in order to ensure that it performs **its intended function**, without failure, for the **required time duration** in a **specified environment** (D. Kiran).

BUTTERWORTH HEINEMANN

Total Quality Management

Key Concepts and Case Studies

2017, Pages 391-404

Reliable machine learning models

Given a new case

And a prediction

Is prediction Reliable?

If not …

- Avoid using prediction for decision making
- If happens «frequently» enough discard model/device

«Key challenger for delivering clinical impact with artificial intelligence», Kelly et al, 2019, BMC Medicine

When a model fails?

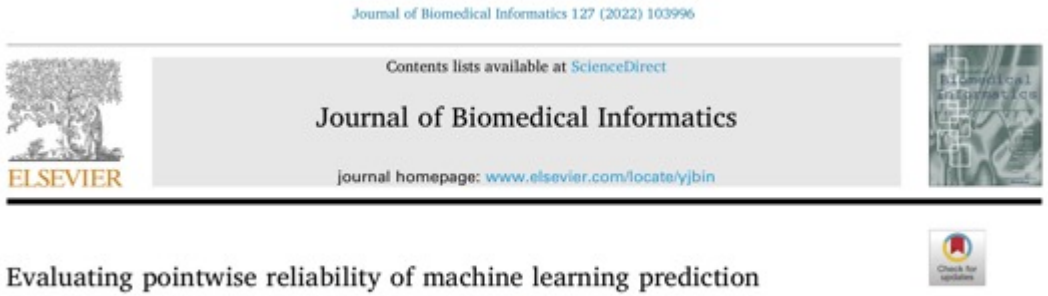
Saria and Subbaswamy,2019, Tutorial: Safe and Reliable Machine Learning

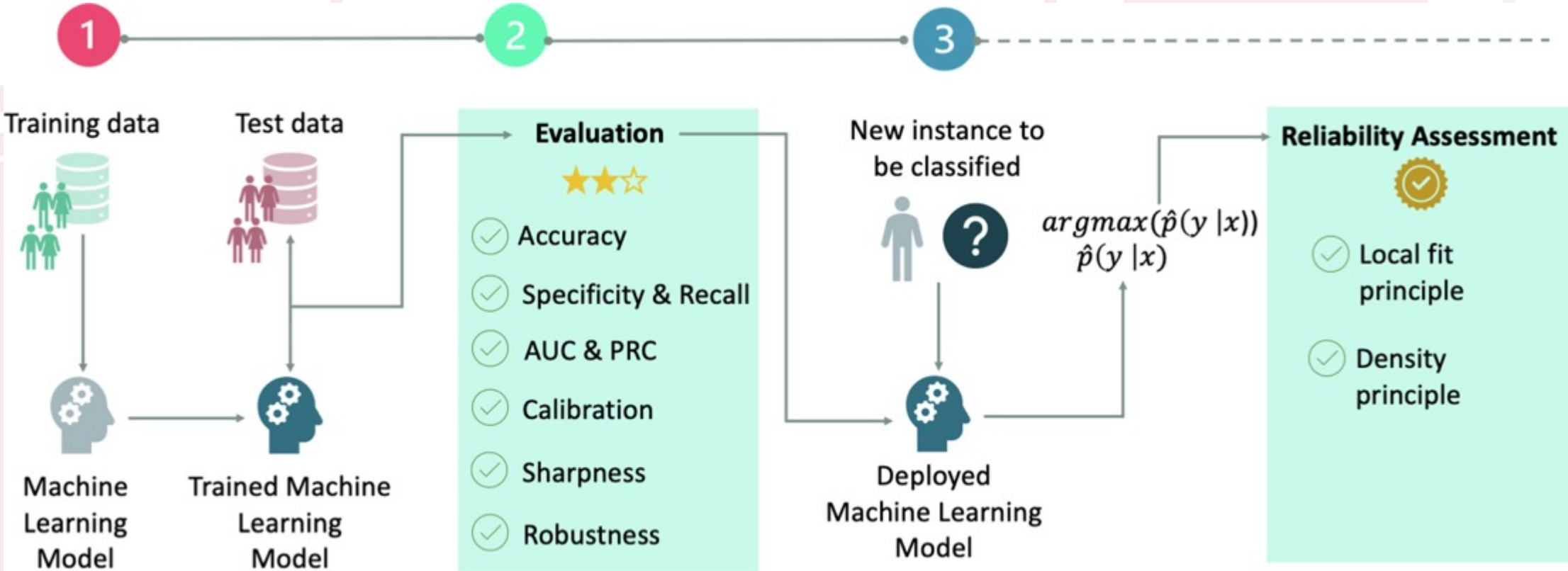
Types of failures:

1. Bad or inadequate data: a particular class or subpopulations are underrepresented, or simply the data do **not** contain **enough information** to solve the problem
2. Differences or **shift in the environment**
3. Model's associated errors: model misspecification, dependent data, **model fragility** (i.e. when the model is applied to high dimensional data, since the prediction is very sensitive to small perturbation in the output)
4. Poor reporting

Reliability

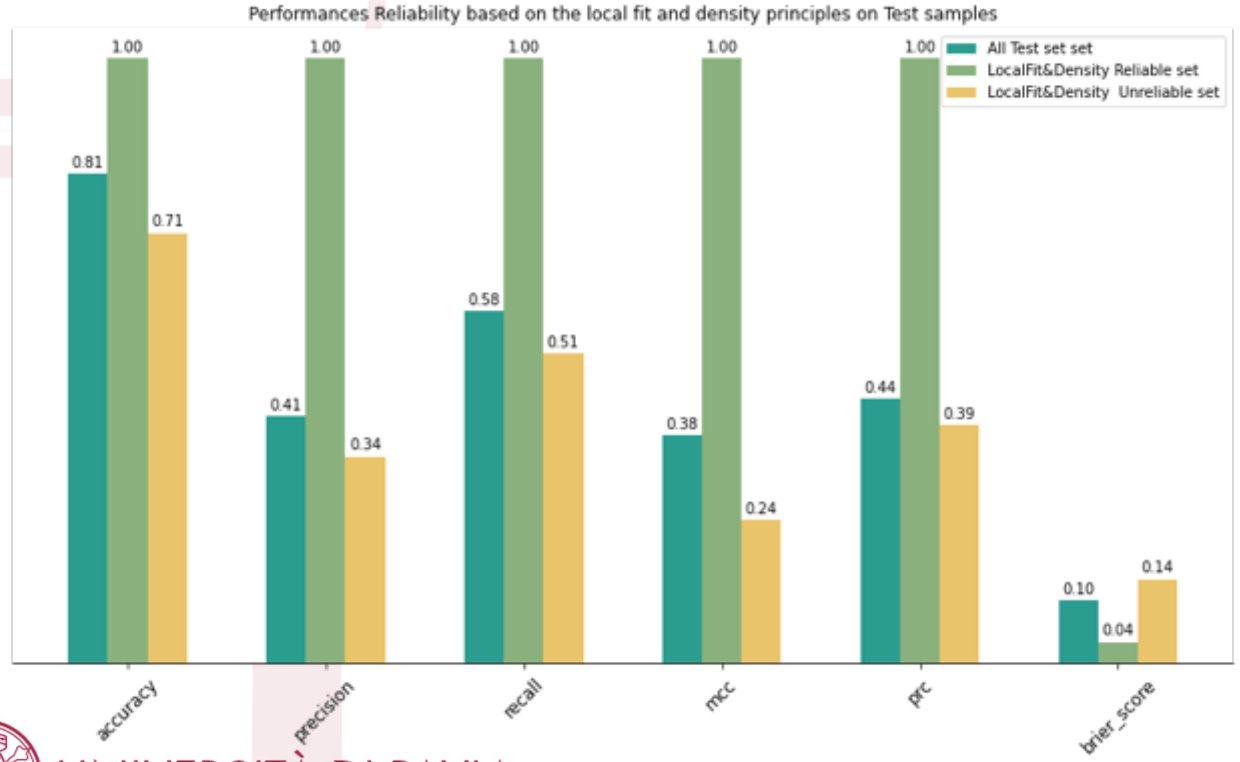
- We use the term “reliability” to denote the degree of reliance on the prediction made by the ML model on a single example.





How to use reliability

- **Selective prediction:** a model can choose to abstain itself from the classification when the reliability is low.



The MIMIC-III dataset & PhysioNet 2012 challenge. Prediction of in-hospital death from clinical data.
4480 patients that survived
768 that died in the hospital.

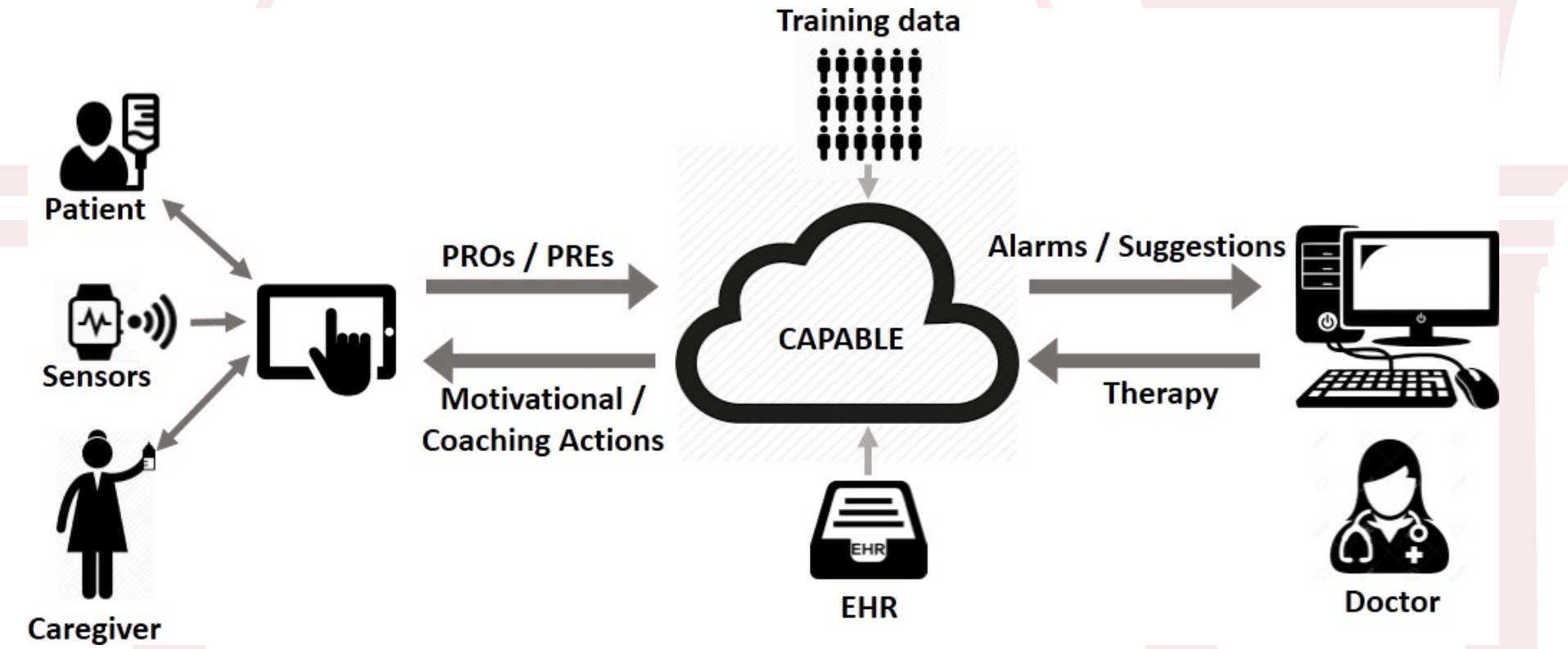
we simulated the extreme case in which only male patients are available in the training set.

Current research directions

- Use a generative model (e.g. autoencoder) to represent training data
- Check local-fit and density principles on the basis of the autoencoder output

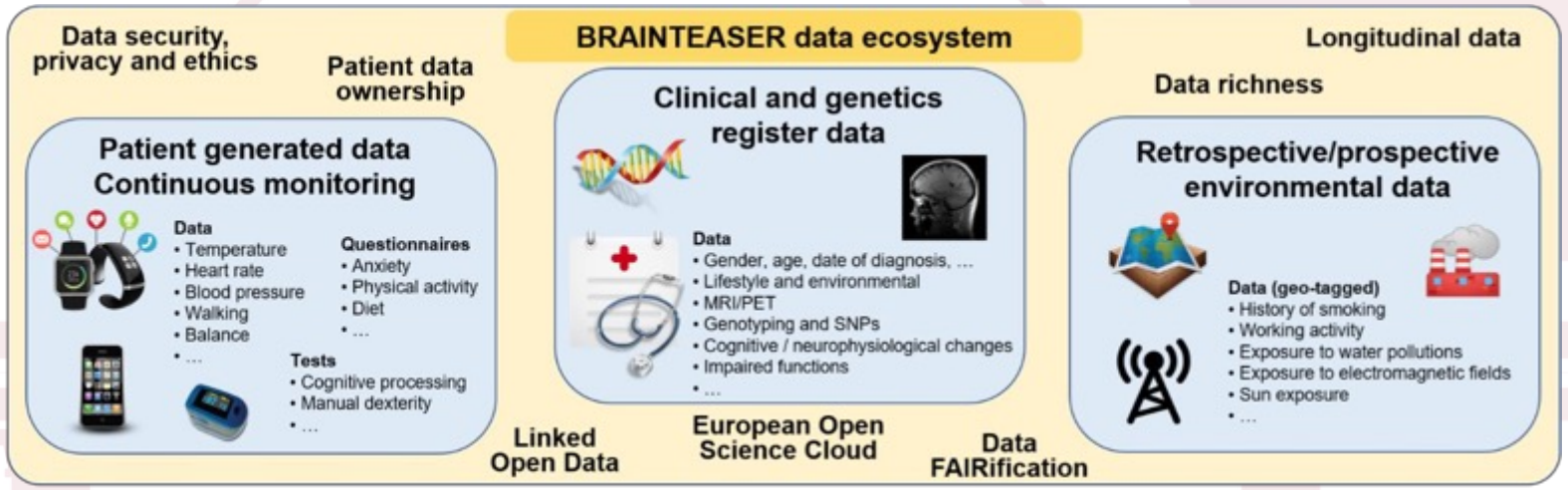
Some projects

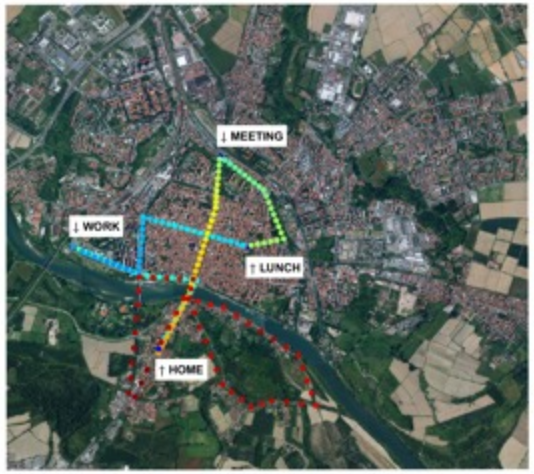
The CAPABLE project
CAncer PAtients: Better Life Experience

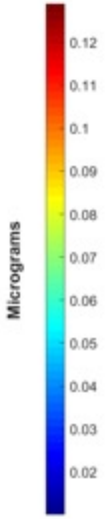


Brainteaser

BRinging Artificial INTelligencE home for a better cAre of **amyotrophic lateral sclerosis and multiple SclERosis**







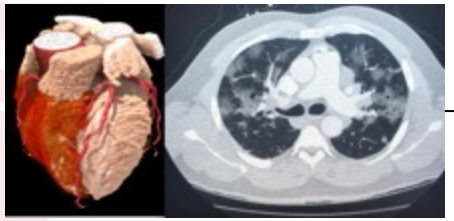
UNIVERSITÀ DI PAVIA

Intestrat-CAD

INTEgrated STRAtification Tools in **Coronary Artery Disease**

Registry Data

CT Scan



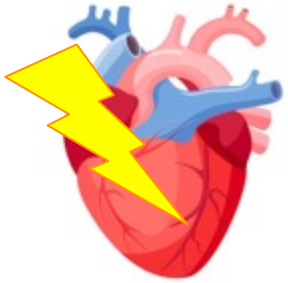
Omics

Genomics and epigenomics

Transcriptomics

Phenotyping

Acute CV Event



PAN-EUROPEAN RESPONSE TO THE IMPACTS OF COVID-19 AND FUTURE PANDEMICS AND EPIDEMICS

periscope

Gather data for COVID - Atlas

Gather data in order to develop a comprehensive, user-friendly, openly accessible COVID Atlas, which should become a reference tool for researchers and policymakers, and a dynamic source of information to disseminate to the general public.

Perform innovative statistical analysis

Perform innovative statistical analysis on the collected data, with the help of various methods, including machine learning tools.

Identify successful practices

Identify successful practices and approaches adopted at the local level, which could be scaled up at the pan-European level for a better containment of the pandemic and its related socio-economic impacts

Develop guidance for policymakers

Develop guidance for policymakers at all levels of government, in order to enhance Europe's preparedness for future similar events and proposed reforms in the multi-level governance of health.

Machine Learning with «moving» data

NIH grant – UFL and UNIPV

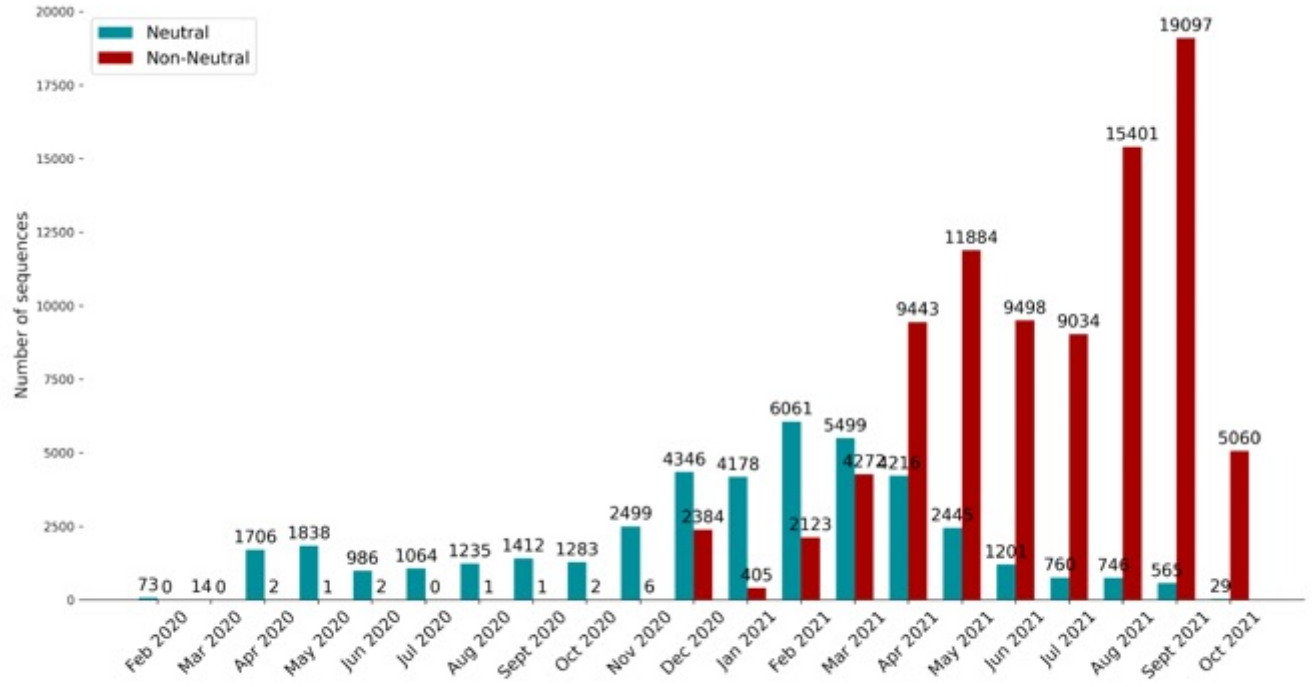


Figure 1. Number of sequences with a percentage of missing values (X) below 1%, submitted to GISAID each month from February, 2020 to October, 2021

Stud Health Technol Inform. 2022 May 25;294:654-656. doi: 10.3233/SHTI220550.

Dynamic Prediction of Non-Neutral SARS-Cov-2 Variants Using Incremental Machine Learning

Giovanna Nicora [1,2], Simone Marini [3], Marco Salemi [4], Riccardo Bellazzi [1]

Affiliations + expand

PMID: 35612170 DOI: 10.3233/SHTI220550

Lessons I have learned

- **Trustworthy AI** - **trust in a socio-technical system**, based on **people** and **ICT + AI technologies**
- AI technological components include a plethora of different solutions, including data-driven, knowledge-driven and mixed approaches
- Data-driven strategies requires:
  - Deep knowledge of the process of care and thus of data generation
  - Transparent protocols for data preprocessing
  - Proper validation strategies
  - Online learning assessment scheme

Lessons I have learned

- There is no single recipe for building trustworthy systems
- Disciplines at the intersection between different fields, such as bioengineering and health informaticians are the “right” communities for properly designing, implementing AND deploying AI solutions
- However:
  - **Trust does not come for free, it requires substantial investments**
  - **Capacity building, AI/MI education are key**

Thanks to …





LISRC

Istituti Clinici Scientifici Maugeri

Giovanna Nicora @ Engenome

SIBIM

UNIVERSITÀ DI PAVIA

Dipartimento di Ingegneria Industriale e dell'Informazione