VITTORIO TORRI

vittorio.torri@polimi.it

Postdoctoral Researcher - Health Analytics

ACADEMIC RESEARCH

During my research experience, I developed a strong foundation in the core statistical, machine learning, and deep learning models commonly applied in **Natural Language Processing**. My research focused on building robust solutions for **scenarios with limited supervision,** where I gained expertise in developing **unsupervised and weakly supervised learning** techniques, as well as **data augmentation** strategies, for tasks traditionally framed as text classification and named entity recognition. In particular, I specialized in processing various types of **medical documents in resource-scarce languages** such as Italian and Dutch, often in combination with structured healthcare data.

Key projects include:

1. Analysis of **Italian referrals**, developing a pipeline for the **extraction of temporal indications** for follow-up examinations, monitoring waiting times and appropriateness of prescriptions - in collaboration with *Regione Lombardia* [P2] [T7]
2. Development of a **clustering pipeline** for **Italian referrals**, monitoring disease spread and appropriateness of prescriptions - in collaboration with *Regione Lombardia* [PP2] [P10] [T7] [T8] [T9] [PS6]
3. Development of a **weakly-supervised classification model** for **Italian discharge letters**, identifying diseases from discharge letters' texts for cohort selection - in collaboration with *University of Padova* [P1] [P6] [P7] [T4] [T7] [PS3] [PS5]
4. **Extraction of structured information** from **Italian clinical notes**, participating in the CLinkaRT shared task of EVALITA 2023 and in the CRF Filling Shared Task of CL4Health 2026 [P8] [PS4] [P3] [PS1]
5. **Extraction of structured information** from **EHR** related to cardiological patients, for their use in subsequent statistical pipelines, exploiting LLMs for data augmentation - in collaboration with *Fondazione Monasterio - Scuola Superiore Sant'Anna Pisa* [P4] [T5] [T7]
6. Development of a model for **interpretable phenotyping of heart failure patients** using both **structured** and **unstructured (textual)** data - in collaboration with *Amsterdam University Medical Center* [PP1] [PS2] [O1]
7. Development of a **language model to predict health trajectories of cardiovascular patients** from heterogeneous and multimodal input sequences - in collaboration with *Human Technopole*
8. *Synthetic data augmentation* for clinical documents using Large Language Models [P5] [T1]

Research interests:

- **Embedding representation** of structured and unstructured patient data for **weakly-supervised and unsupervised learning**
- **Language models** for clinical documents in under-resourced languages
- **Interpretability** of black-box language models
- Generation and evaluation of structured and unstructured **synthetic** data

WORK EXPERIENCE

Health Analytics group, MOX laboratory, Department of Mathematics, Politecnico di Milano *May 2025 -*
Postdoctoral Researcher
Supervisor: prof. Francesca Ieva
Research focused on the development of statistical and machine learning models that can exploit comprehensive semantic representation of patients from multiple data sources to extract actionable insights for healthcare providers

Health Data Science Centre, Fondazione Human Technopole *Dec 2024 -*

Visiting Researcher

Supervisor: prof. Francesca Ieva

Visiting researcher working on a project related to the use of LLMs to predict health trajectories of cardiovascular patients

EDUCATION

Health Analytics group, MOX laboratory, Department of Mathematics, Politecnico di Milano *Feb 2022 - May 2025*

PhD in Data Analytics and Decision Sciences cum laude

Thesis: *Natural Language Processing for clinical documents: Addressing Label Scarcity to Support Healthcare Policies and Clinical Practice*

Supervisor: prof. Francesca Ieva

Research focused on the investigation and the development of statistical and machine learning models to extract value from textual data in the healthcare domain, in collaboration with public health authorities from Regione Lombardia and multiple healthcare providers.

Amsterdam University Medical Center *Oct 2023 - March 2024*

Visiting PhD Student

Department of Medical Informatics & Department of Cardiology

Supervisors: dr. M.J.Boonstra, dr. I. Calixto, prof. A. Abu-Hanna, prof. F.W. Asselbergs

Visiting PhD student working on the development of a model for interpretable phenotyping heart failure patients combining structured and textual clinical data

Politecnico di Milano *Sep 2019 - Dec 2021*

Master of Science in Computer Science and Engineering - 110/110 cum laude

Thesis: *Textual eXplanations for intuitive Machine Learning*

Supervisor: prof. Mark J. Carman

Thesis on the development and evaluation of a generative language model for textual explanations of black-box machine learning classifiers

Main courses: *Machine Learning, Artificial Neural Networks, Data Mining, Data Intelligence Applications*

Politecnico di Milano *Sep 2016 - Sep 2019*

Bachelor of Science in Computer Science and Engineering - 110/110 cum laude

Main Courses: *Fundamentals of Computer Science, Mathematical Analysis, Probability & Statistics, Algorithms and Data Structures, Computer Architectures and Operating Systems, Software Engineering*

PUBLICATIONS AND CONFERENCES

Publications

[P1] *V. Torri, E. Barbieri, A. Cantarutti, C. Giaquinto, F. Ieva* - Automatic identification of diagnosis from hospital discharge letters via weakly supervised Natural Language Processing. -Scientific Reports, 16, 26727, 2026.

[P2] *V. Torri, M. Ercolanoni, F. Bortolan, O. Leoni, F. Ieva* - A NLP-based semi-automatic identification system for delays in follow-up examinations: an Italian case study on clinical referrals - BMC Medical Informatics and Decision Making 24.1: 107, 2024.

[P3] *V. Torri, F. Ieva* - Polimi at CRF Filling 2026: Prompt-Based Information Extraction from Italian Clinical Notes - In Proceedings of the Third Workshop on Patient-Oriented Language Processing (CL4Health) @ LREC 2026, 412-427, ELRA, 2026.

[P4] *V. Torri, S. Mazzucato, S. Dalmiani, U. Paradossi, C. Passino, F. Ieva* - Structuring Clinical Notes of Italian ST-elevation Myocardial Infarction Patients - Proceedings of the First Workshop on Patient-Oriented Language Processing (CL4Health)@ LREC-COLING 2024 - ELRA and ICCL, 2024.

[P5] *V. Torri, F. Ieva* - Analyzing the Impact of LLM-Based Augmentation on Document Classification. - In JADT 2026 Proceedings Volume I, 242-251, Palermo University Press, 2026.

[P6] *V. Rigamonti, V. Torri, S. Morris, D. Donà, F. Ieva, C. Giaquinto, C. Di Chiara, A. Cantarutti* - Live Attenuated vs. Inactivated Influenza Vaccines: a Population-based Assessment of Effectiveness in Italian Children - Vaccine, 53, 126946, 2025.

[P7] *V. Rigamonti, A. Cantarutti, V. Torri, F. Ieva, C. Giaquinto, C. Di Chiara, D. Donà, CARICE study group* - Real-world comparison of live attenuated and inactivated influenza vaccine effectiveness in Italian children - Vaccine: X, 31, 100856, 2026.

[P8] *V. Torri, F. Ieva* - Polimi at CLinkaRT: a Conditional Random Field vs a BERT-based approach - Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023) - CEUR. org, 2023

[P9] *V. Torri, F. Ieva* - Graph Neural Networks for clustering medical documents - Proceedings of the Statistics and Data Science 2024 Conference - New perspectives on Statistics and Data Science - Palermo University Press, 2024

[P10] *V. Torri, M. Ercolanoni, F. Bortolan, O. Leoni, F. Ieva* - Clustering Italian medical texts: a case study on referrals - Proceedings of the Statistics and Data Science Conference - Pavia University Press, 2023.

[P11] *V. Torri, F. Ieva* - Leveraging LLMs for medical document clustering: Towards a benchmark dataset - In Statistical Methods for Data Analysis and Decision Sciences, Italian Statistical Society Series on Advances in Statistics. Springer, To Appear (2026).

[P12] *L. Mazzeo, F. Corso, P. Baili, F. Scotti, V. Torri, et al.* - Data analytics for real-world data integration in TKI-treated NSCLC patients using electronic health records - ESMO Real World Data and Digital Oncology, 7:100109, 2025.

[P13] *F. Corso, P. Baili, F. Scotti, L. Mazzeo, V. Torri, et al.* - Cost-effectiveness of first-line osimertinib informed by electronic medical records via text-mining: a real-world Italian case study of EGFR-mutated advanced NSCLC patients - ESMO Real World Data and Digital Oncology, 10:100198, 2025.

Pre-prints and Submitted Works

[PP1] *V. Torri, M. J. Boonstra, M. C. van de Veerdonk, D. N. Kalkman, A. Uijl, F. Ieva, A. Abu-Hanna, F. W. Asselbergs, I. Calixto* - NLP-based assessment of prescription appropriateness from Italian referrals - arXiv preprint arXiv:2505.24619 , 2025 - *Submitted*.

[PP2] *V. Torri, A. Bottelli, M. Ercolanoni, O. Leoni, F. Ieva* - NLP-based assessment of prescription appropriateness from Italian referrals - arXiv preprint arXiv:2501.14701 , 2025 - *Submitted*.

Talks

[T1] *V. Torri, F. Ieva* - Investigating Class-Level Variability in Synthetic Data Augmentation for Document Classification - at the *18th International Conference on Statistical Analysis of Textual Data - JADT 2026* - Palermo, 8 July 2026

[T2] *V. Torri, F. Ieva* - Investigating Class-Level Variability in Synthetic Data Augmentation for Document Classification - at *Statistics and Data Science Conference 2026* - Caserta, 1 April 2026

[T3] *V. Torri, F. Ieva* - Leveraging LLMs for Medical Document Clustering: Towards a Benchmark Dataset - Invited at *Statistics and Data Science Conference 2025* - Milano, 2 April 2025

[T4] *V. Torri, E. Barbieri, A. Cantarutti, C. Giaquinto, F. Ieva* - Weakly-supervised classification of clinical documents: a case study on Italian discharge letters - at *Precision Health Day 2025* - Lugano, 28 March 2025

[T5] *V. Torri, S. Mazzucato, S. Dalmiani, U. Paradossi, C. Passino, F. Ieva* - Structuring Clinical Notes of Italian ST-elevation Myocardial Infarction Patients - at *CL4Health - First Workshop on Patient-Oriented Natural Language Processing @ LREC-COLING 2024* - Torino, 20 May 2024

[T6] *V. Torri, F. Ieva* - Graph Neural Networks for clustering medical documents - at *Statistics and Data Science Conference 2024* - Palermo, 12 April 2024

[T7] *V. Torri, F. Ieva* - Natural Language Processing for Mining Healthcare Utilization Databases - Invited at *The Role of Statistics in Medicine and Healthcare Research Workshop* - Roma, 15 March 2024

[T8] *V. Torri, M. Ercolanoni, F. Bortolan, O. Leoni, F. Ieva* - Natural Language Processing to support health policies: a case study on referrals - at *1st Workshop on Advanced Quantitative Methods and Analytics for Public Policy Support* - Milano, 26 October 2023

[T9] *V. Torri, M. Ercolanoni, F. Bortolan, O. Leoni, F. Ieva* - Clustering Italian medical texts: a case study on referrals - at *Statistics and Data Science Conference 2023* - Pavia, 27 April 2023

Poster presentations

[PS1] *V. Torri, F. Ieva* - Polimi at CRF Filling 2026: Prompt-Based Information Extraction from Italian Clinical Notes - at *Third Workshop on Patient-Oriented Language Processing (CL4Health) @ LREC 2026* - Palma, 12 May 2026

[PS2] [Best Poster Award] *V. Torri, M. J. Boonstra, M. C. van de Veerdonk, D. N. Kalkman, A. Uijl, F. Ieva, A. Abu-Hanna, F. W. Asselbergs, I. Calixto* - NLP-based assessment of prescription appropriateness from Italian referrals - at *Precision Health Day 2026* - Lugano, 27 March 2026

[PS3] *V. Torri, E. Barbieri, A. Cantarutti, C. Giaquinto, F. Ieva* - Weakly-supervised classification of clinical documents: a case study on Italian discharge letters - at *Mathematics for our Health Workshop 2024* (M4H) - Milano, 7-8 November 2024

[PS4] *V. Torri, F. Ieva* - Polimi at CLinkaRT: a Conditional Random Field vs a BERT-based approach - at *Final Workshop of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian* (EVALITA 2023) - Parma, 7-8 September 2023

[PS5] *V. Torri, E. Barbieri, A. Cantarutti, C. Giaquinto, F. Ieva* - Weakly-supervised classification of clinical documents: a case study on Italian discharge letters - at *44th International Society for Clinical Biostatics Conference* (ISCB44) - Milano, 27-31 August 2023

[PS6] *V. Torri, M. Ercolanoni, F. Bortolan, O. Leoni, F. Ieva* - Natural language processing for Italian medical documents: a case study on referrals clustering - at *Eighth Phd Meeting Workshop* organized by Istituto di Ricerche Farmacologiche Mario Negri - Milano, 29-30 June 2023

Organizing Committee

Member of the Local Organizing Committee of the *1st Workshop on Data Science for Health and Biology (DS4HB)* held in Milano on 7-9 April 2025

Other

[O1] Invited Seminar at the Department of Medical Informatics, Amsterdam UMC - Interpretable phenotyping of Heart Failure patients with Dutch discharge letters, 17 October 2024

[O2] Invited Divulgative Seminar to the citizenship of La Spezia, organized by ASL 5 La Spezia - *I dati sanitari spiegati ai cittadini (Healthcare data explained to citizens)*, 16 May 2024

[O3] Lecture on *Natural Language Processing* as part of the course *Advanced Statistical Learning for Complex Data* for PhD Students in Data Analytics and Decision Sciences at Politecnico di Milano, 15 May 2024 and 3 February 2026

TEACHING EXPERIENCE

Politecnico di Milano *Spring 2025, Spring 2026*
Teaching Assistant
Esercitatore (TA) for the course *Applied Statistics* (5 ECTS - 1st year MSc High Performance Computing, Computer Science and Engineering, Physics Engineering - 20h/semester) - lecturer dr. Michela Carlotta Massi (2026), dr. Mario Beraha (2025)

Politecnico di Milano, Università Vita-Salute San Raffaele *Fall 2024, Fall 2025*
Teaching Assistant
Esercitatore (TA) for the course *Statistical Models for Healthcare Data* (6 ECTS - 1st year MSc Health Informatics - 24h/semester) - lecturer dr. Lara Cavinato

Politecnico di Milano *Spring 2022, Spring 2023*
Teaching Assistant

Esercitatore (TA) for the course *Statistica* (Statistics) (5 ECTS - 1st year BSc in Mathematical Engineering - 24h/semester) - lecturer prof. Simone Vantini

Politecnico di Milano *Fall 2022, Spring 2024*

Tutor

Tutor for *Statistica e Calcolo delle Probabilità* (Statistics and Probability Calculus) (20h/semester) for all bachelor students in engineering

Politecnico di Milano *Spring 2024, Spring 2025, Spring 2026*

Tutor

Tutor for three group projects in the course *Applied Statistics* (MSc in Mathematical Engineering, MSc in Computer Science and Engineering)

Politecnico di Milano *Sep 2020 - Dec 2021*

Tutor

Tutor for multiple Computer Science courses of various bachelors in engineering (*Informatica A, Informatica B, Fondamenti di Informatica, Basi di Dati, Progetto di Algoritmi e Strutture Dati*) covering basic and advanced C programming, SQL and relational database design and basic MATLAB programming.

Politecnico di Milano

Master Thesis Co-Supervisor

- Master Thesis in Mathematical Engineering - Statistical learning track - *"Combination of text mining and statistical learning techniques for the generation of real-world evidence in oncology: a case study for patients affected by NSCLC with EGFR mutation"* - Author: Anna Iob - Supervisor: dr. Andrea Cappozzo (Jul - Dec 2022)
- Master Thesis in Mathematical Engineering - Statistical learning track - *"Graph Neural Networks for clustering medical documents"* - Author: Gabriele Moro - Supervisor: prof. Francesca Ieva (Sep 2023 - Jul 2024)
- Master Thesis in Computer Science and Engineering - *"Prediction and Learning of Disease Paths Leading to Death via Generative Models"* - Author: Khalid Fathelrahman - Supervisor: prof. Francesca Ieva (Dec 2024 - Jul 2025)
- Master Thesis in Computer Science and Engineering - *"A Methodological Framework for Evaluating Synthetic Longitudinal Clinical Aata: a Case Study on Heart Failure Patients in Lombardy Region"* - Author: Manuela Marenghi - Supervisor: prof. Francesca Ieva (Sept 2025 - Jul 2026)

SKILLS

Languages: Italian - mothertongue
English - good (TOEIC Listening & Reading 990/990 - Aug 2019)

Technical: Good knowledge of **Python**, in particular of the main data science libraries, including deep-learning libraries (e.g.: *pandas*, *scikit-learn*, *keras*, *pytorch*)
and libraries for NLP (e.g.: *transformers*, *vllm*, *gensim*, *transformers*, *nltk*)
Good knowledge of **R**
Good knowledge of **SQL** and relational databases
Basic knowledge of **non-relational** databases
Good knowledge of **C**, basic knowledge of **C++**
Discrete knowledge of **Java**
Basic knowledge of **MATLAB**
Basic knowledge of web languages **(HTML, CSS, JS, PHP)**
Git, **Latex**, **Microsoft Office** suite, **Windows** and **Linux** OS

Soft skills: Through my academic studies, research projects, and university teaching experience, I have developed strong **teamworking** skills.

Teaching activity, including years of private lessons, interactions with institutions such as Regione Lombardia and multiple healthcare providers, and conference presentations, enhanced my **communication** skills.

Managing complex research timelines and academic commitments throughout my PhD strengthened my **organizational** and **time-management** skills.

Autorizzo al trattamento dati ai sensi del GDPR 2016/679 del 27 aprile 2016 (Regolamento Europeo relativo alla protezione delle persone fisiche per quanto riguarda il trattamento dei dati personali).

23 September 2026