Verifiable and Reproducible Clinical Registry Form Completion from Patient Documents

Petr Nazarov

Svetlana Simova (1), Petr V. Nazarov (1), Sophie Servais (2,3), Sébastien Rinaldetti (1,2), José Batista da Costa (4), Vladimir Despotovic (1), Maximilian Fünfgeld (1)
1) Luxembourg Institute of Health, Strassen
2) Centre Hospitalier de Luxembourg, Luxembourg
3) University of Luxembourg, Esch-sur-Alzette
4) Les Hôpitaux Robert Schuman, Luxembourg

Completing clinical registry case report forms (CRFs) from heterogeneous patient records remains labor-intensive. Although large language models (LLMs) can accelerate clinical information extraction, unsupported outputs, limited reproducibility, and data-protection requirements remain major barriers to clinical deployment [1].

We developed MedForms Assistant, a privacy-preserving system for automated CRF completion in which every generated value is reviewable, traceable, and reproducible. The system combines locally deployed LLMs with deterministic processing in a pipeline comprising document ingestion, OCR, classification, content-specific processing, AI-assisted extraction, evidence verification, rule-based derivation, form rendering, and audit logging. The workflow can be adapted to different CRFs while keeping patient documents within the protected network.

The prototype was implemented for the EBMT Acute Leukaemias Diagnosis CRF, using acute myeloid leukemia as the first use case. Of 118 fields, 47 are queried from a local LLM (gpt-oss-20b), while 71 are derived deterministically from extracted information. Each populated value must be linked to the supporting text in the source document. This evidence-first design follows work emphasizing mechanically verifiable support for biomedical LLM outputs [2,3]. Each run produces completed predefined forms (PDF, DOCX, XLSX, TXT) and a provenance-rich JSON file without requesting Internet resources. The system is designed as a human-in-the-loop workflow in which generated results require clinician review before use. We are currently adapting the tool for DKG/ECC prostate cancer center certification documentation.

Our tool provides a practical architecture for secure, verifiable, and reproducible clinical documentation automation. Key strengths are traceability, deterministic derivation, local deployment, and explicit human oversight; limitations include dependence on local computational resources and the absence of an independent clinical accuracy benchmark – the work is ongoing.

References

1. Ferrazzi P, et al. Overview of the CRF 2026 Shared Task on Clinical Case Report Forms Filling. CL4Health @ LREC 2026. 2026:245–254.

2. Windisch P, et al. Verbatim Evidence Requirements and Automated Assessment of Large Language Models for Biomedical Text Processing of Trial Eligibility Criteria. Cureus. 2026;18(5):e108666.

3. Yapo J, et al. Combining Anti-Hallucination Strategies for Reliable LLM-Based Clinical Information Extraction. Studies in Health Technology and Informatics. 2026;336:1062–1063.

Biosketch

Petr NAZAROV, PhD

Head of Bioinformatics and AI (BIOAI) Unit, Department of Medical Informatics, Luxembourg Institute of Health (LIH)
ORCID ID: 0000-0003-3443-0298

CURRENT AND PAST POSITIONS

2024–present: Head of Bioinformatics and AI (BIOAI) Unit, Department of Medical Informatics, Luxembourg Institute of Health (LIH).

2019–present: Group Leader of Multiomics Data Science (MODAS) Research Group, Department of Cancer Research, LIH.

2010–present: Associate Assistant Professor, University of Luxembourg.

2020–2023: ad interim Head of Bioinformatics Platform, LIH.

2007–2019: Researcher in biostatistics, genomics and bioinformatics, LIH (former CRP-Santé).

2006–2007: Postdoctoral researcher (FNR Inter-Mobility fellowship), CRP-Santé, Luxembourg.

2001–2007: Engineer-programmer, then Assistant Professor, Belarusian State University, Minsk.

EDUCATION

2021: ADR (autorisation à diriger des recherches), University of Luxembourg.

2006: PhD in Biophysics, Wageningen University, the Netherlands.

2001: MSc in Systems Analysis and Modelling, Belarusian State University, Belarus.

AWARDS AND HONORS

Grants held as PI, co-PI or work-package leader: MARKOPOLO – markers of pollution (Horizon Europe, 2025-2028, WP leader); PEAR-OP – omics profiling by Nanopore (LIH Intrepid, 2024-2026, co-PI); DIOMEDES – glioblastoma ecosystem (FNR CORE, 2022–2025, co-PI); SUNRISE – DNA damage in glioma (Télévie, 2022–2026, co-promoter); RADIO-D2R2 – radiomics for radio-necrosis vs. progression (Télévie, 2024–2026, co-promoter); Foundation-METS – foundation model for brain oligometastases (Télévie, 2025–2027, co-promoter); PERSIST – glioblastoma persister cells and the immune microenvironment (Télévie, 2025–2027, co-promoter); DEMICS – mixed transcriptome decomposition (FNR CORE, 2018–2021, PI); ActinSim – actin motility modelling (FNR Mobility, 2006–2007, PI).

Partner or co-investigator in CLINNOVA, EATRIS-Plus, COVIRNA, GLASS-LUX and CANBIO/CANBIO2 (2018–2030).

OTHER RELEVANT PROFESSIONAL ACTIVITIES AND ACCOMPLISHMENTS

Publications: over 90 peer-reviewed publications; h-index 32 (Google Scholar).

Grant review panels: INSERM, France (2025, 2026); Israel Science Foundation (2026); Fondation Pélican, Luxembourg (2024–2026); Novo Nordisk Fonden, Denmark (2023).

Open-source software: consICA and ORFhunteR (Bioconductor), CoExpress; contributions to DecompPipeline/MeDeCom.

Supervision (2016–2026): 4 postdoctoral researchers, 2 PhD students, 6 MSc students; member of 4 PhD defence committees.

Teaching: PhD course “Multiomics Data Science” (LIH, since 2020); BSc – Advanced Biostatistics, University of Luxembourg (2012-2024) and other. Materials at edu.modas.lu.

Member: European Association for Cancer Research; International Society for Computational Biology.