Medical Information Extraction with NLP-Powered QABots: a Real-World Scenario

Claudio Crema, Federico Verde, Pietro Tiraboschi, Camillo Marra, Andrea Arighi, Silvia Fostinelli, Guido Maria Giuffré, Vera Pacoova Dal Maschio, Federica L'Abbate, Federica Solca, Barbara Poletti, Vincenzo Silani, Emanuela Rotondo, Vittoria Borracci, Roberto Vimercati, Valeria Crepaldi, Emanuela Inguscio, Massimo Filippi, Francesca Caso, Alessandra Maria Rosati, Davide Quaranta, Giuliano Binetti, Ilaria Pagnoni, Manuela Morreale, Francesca Burgio, Michelangelo Stanzani Maserati, Sabina Capellari, Matteo Pardini, Nicola Girtler, Federica Piras, Fabrizio Piras, Stefania Lalli, Elena Perdixi, Gemma Lombardi, Sonia Di Tella, Alfredo Costa, Marco Capelli, Cira Fundarò, Marina Manera, Cristina Muscio, Elisa Pellencin, Raffaele Lodi, Fabrizio Tagliavini, Alberto Redolfi

This work was funded in part by: the National funding of Italian Ministry of Economy and Finance (CCR-2017-23669078), the National funding of the Italian Ministry of Health under the frame-work of the grant ISTITUTI NAZIONALI VIRTUALI (RCR 2020–23670067 and RCR-2021-23671214), in the frame-work of the grant PROGETTO RETE RIN 2022 (RCR-2022-23682294), and by the Ministry of Health under the IRCCS Research Program - Ricerca Corrente 2023-2024, Linea n. 2 "Piattaforme elettroniche per analisi di immagini cerebrali".

(Corresponding author: C. Crema).

Claudio Crema and Alberto Redolfi are with the Laboratory of Neuroinformatics (e-mail: ccrema@fatebenefratelli.eu. aredolfi@fatebenefratelli.eu), Silvia Fostinelli and Giuliano Binetti are with the MAC - Memory Clinic and Molecular Markers Laboratory (e-mail: sfostinelli@fatebenefratelli.eu, gbinetti@fatebenefratelli.eu), and Ilaria Pagnoni is with the Neuropsychology Unit (e-mail: ipagnoni@fatebenefratelli.eu), all afferent to the IRCCS Istituto Centro San Giovanni di Dio Fatebenefratelli, Brescia, Italy.

Federico Verde, Federica Solca, Barbara Poletti, and Vincenzo Silani are with the Department of Neurology and Laboratory of Neuroscience of the IRCCS Istituto Auxologico Italiano, Milan, Italy. F. V. and V. S. are also with the Department of Pathophysiology and Transplantation of the Dino Ferrari Centre, Università degli Studi di Milano, Milan, Italy. B. P. is also with the Department of Oncology and Hemato-Oncology of the Università degli Studi di Milano, Milan, Italy (e-mail: f.verde@auxologico.it, f.solca@auxologico.it, b.poletti@auxologico.it, vincenzo@silani.com).

Pietro Tiraboschi, Valeria Crepaldi, Emanuela Inguscio, and Elisa Pellencin are with the Division of Neurology (e-mail: pietro.tiraboschi@istituto-besta.it, valeria.crepaldi@istituto-besta.it, emanuela.inguscio@istituto-besta.it, elisa.pellencin@istituto-besta.it), and Fabrizio Tagliavini is with the Scientific Directorate, all afferent to the Fondazione IRCCS Istituto Neurologico Carlo Besta, Milan, Italy (e-mail: fabrizio.tagliavini@istituto-besta.it). Cristina Muscio was with the Division of Neurology of the Fondazione IRCCS Istituto Neurologico Carlo Besta, Milan, Italy. She is now with the ASST Bergamo Ovest, Bergamo, Italy (e-mail: cristina_muscio@asst-bgovest.it).

Camillo Marra, Guido Maria Giuffré, Federica L'Abbate, and Alessandra Maria Rosati are with the with the Memory Clinic of the IRCCS Policlinico A. Gemelli Foundation. Davide Quaranta is with the Neurology Unit of the IRCCS Policlinico A. Gemelli Foundation (e-mail: camillo.marra@policlinicogemelli.it, guido.giuffre@gmail.com, federicalabbate@gmail.com, alessandramaria.rosati@guest.policlinicogemelli.it, davide.quaranta@policlinicogemelli.it).

Andrea Arighi, Emanuela Rotondo, Vittoria Borracci, and Roberto Vimercati are with the Neurodegenerative Diseases Unit of the Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico, Milan, Italy (e-mail: andrea.arighi@policlinico.mi.it, emanuela.rotondo@policlinico.mi.it, vittoria.borracci@policlinico.mi.it, roberto.vimercati@policlinico.mi.it).

Vera Pacoova Dal Maschio is with the Department of Neuroscience "Rita Levi Montalcini" of the University of Torino, Torino, Italy, and with the Neurology 2 Unit, A.O.U. Città della Salute e della Scienza di Torino, Torino, Italy (e-mail: vera.pacoovadalmaschio@unito.it).

Massimo Filippi and Francesca Caso are with the Neurology Unit of the IRCCS Ospedale San Raffaele, Milano, Italy. M. F. is also with the Neurorehabilitation Unit, the Neurophysiology Service, and Neuroimaging Research Unit, Division of Neuroscience, all afferent to the IRCCS Ospedale San Raffaele, Milano, Italy. He is also with the Vita-Salute San Raffaele University, Milano, Italy (e-mail: filippi.massimo@hsr.it, caso.francesca@hsr.it).

Manuela Morreale is with the Oasi Research Institute-IRCCS, Troina, Italy (e-mail: mmorreale@oasi.en.it).

Francesca Burgio is with the Neuropsychology Department of the IRCCS San Camillo Hospital, Venice, Italy (e-mail: francesca.burgio@hsancamillo.it).

Michelangelo Stanzani Maserati, Sabina Capellari and Raffaele Lodi are with the IRCCS Istituto delle Scienze Neurologiche di Bologna, Italy. S. C. and R. L. are also with the Department of Biomedical and Neuromotor Sciences of the University of Bologna, Italy (e-mail: m.stanzanimaserati@ausl.bologna.it, sabina.capellari@unibo.it, raffaele.lodi@unibo.it).

Matteo Pardini and Nicola Girtler are with the Department of Neuroscience, Rehabilitation, Ophthalmology, Genetics, Maternal and Child Health (DINOGMI) of the University of Genoa, Genoa, Italy, and also with the IRCCS Ospedale Policlinico S. Martino, Genoa, Italy (e-mail: matteo.pardini@unige.it, nicolagirtler@unige.it).

Federica Piras and Fabrizio Piras are with the Clinical Neuroscience and Neurorehabilitation Department of the IRCCS Santa Lucia Foundation, Rome, Italy (e-mail: federica.piras@hsantalucia.it, f.piras@hsantalucia.it).

Stefania Lalli and Elena Perdixi are with the Department of Neurology of the IRCCS Humanitas Research Hospital, Milan, Italy (e-mail: stefania.lalli@humanitas.it, elena.perdixi@humanitas.it).

Gemma Lombardi is with the Department of Neurosciences, Psychology, Drug Research and Child Health (NEUROFARBA) of the University of Florence, Italy, and also with the IRCCS Fondazione Don Carlo Gnocchi ONLUS Florence, Italy (e-mail: glombardi@dongnocchi.it). Sonia Di Tella is with the Department of Psychology of the Università Cattolica del Sacro Cuore, Italy, and also with the IRCCS Fondazione Don Carlo Gnocchi ONLUS Milan, Italy (e-mail: sditella@dongnocchi.it).

Alfredo Costa and Marco Capelli are with the Unit of Behavioral Neurology of the IRCCS Mondino Foundation Pavia, Italy. A. C. is also with the Department of Brain and Behavioral Sciences of the University of Pavia, Pavia, Italy (e-mail: alfredo.costa@mondino.it, marco.capelli@mondino.it).

Cira Fundarò is with the Neurophysiopatology Unit of the IRCCS Istituti Clinici Scientifici Maugeri, Pavia, Italy (e-mail: cira.fundaro@icsmaugeri.it). Marina Manera is with the Psychology Unit of the IRCCS Istituti Clinici Scientifici Maugeri, Pavia, Italy (e-mail: marina.manera@icsmaugeri.it).

***Abstract*— The advent of computerized medical recording systems in healthcare facilities has made data retrieval tasks easier, compared to manual recording. Nevertheless, the potential of the information contained within medical records remains largely untapped, mostly due to the time and effort required to extract data from unstructured documents. Natural Language Processing (NLP) represents a promising solution to this challenge, as it enables the use of automated text-mining tools for clinical practitioners. In this work, we present the architecture of the Virtual Dementia Institute (IVD), a consortium of sixteen Italian hospitals, using the NLP Extraction and Management Tool (NEMT), a (semi-) automated end-to-end pipeline that extracts relevant information from clinical documents and stores it in a centralized REDCap database. After defining a common Case Report Form (CRF) across the IVD hospitals, we implemented NEMT, the core of which is a Question Answering Bot (QABot) based on a modern NLP model. This QABot is fine-tuned on thousands of examples from IVD centers. Detailed descriptions of the process to define a common minimum dataset, Inter-Annotator Agreement calculated on clinical documents, and NEMT results are provided. The best QABot performance show an Exact Match score (EM) of 78.1%, a F1-score of 84.7%, a Lenient Accuracy (LAcc) of 0.834, and a Mean Reciprocal Rank (MRR) of 0.810. EM and F1 scores outperform the same metrics obtained with ChatGPTv3.5 (68.9% and 52.5%, respectively). With NEMT the IVD has been able to populate a database that will contain data from thousands of Italian patients, all screened with the same procedure. NEMT represents an efficient tool that paves the way for medical information extraction and exploitation for new research studies.**

***Index Terms*— Natural language processing, Question answering (information retrieval), Text mining, Biomedical informatics, Clinical neuroscience.**

I. INTRODUCTION

Digital technologies are becoming pervasive in healthcare facilities [1], particularly in Scientific Institutes for Research, Hospitalization and Healthcare (named IRCCS, from the Italian acronym) in which care and research are combined, thus entailing the prompt availability of large amounts of data. These technologies are leading to a significant increase of digitized textual medical data in the everyday clinicians practice (e.g., discharge letters, exams results, medical notes) [2]. These documents, usually referred to as electronic Case Report Forms (eCRFs), are extremely informative, albeit their creation is time-consuming: medical practitioners spend around 35% of their working time on this activity [3]. Moreover, due to their unstructured nature, these data are often not fully utilized to answer medical questions, which impairs the efficiency of the clinical setting [43]. For these reasons, tools capable of extracting data of interest from unstructured eCRFs and to store them in organized databases (DBs) could greatly increase both the efficiency and efficacy of clinical routines, also easing the process of medical records data-mining to orient research questions. This article, in the context of the Italian Neuroscience and Rehabilitation Network (RIN, Rete IRCCS delle Neuroscienze e della Neuroriabilitazione, https://www.reteneuroscienze.it/en/), describes the implementation process of a semi-automatic pipeline that extracts clinical data from eCRFs and performs data-entry into a centralized DB. RIN is composed by the Virtual National Institutes (IVN, Istituti Virtuali Nazionali). They harmonize IRCCSs activities, rationalize investments and resources, build large cohorts, and interact with international networks [4]. The first established IVN was the Virtual Dementia Institute (IVD, Istituto Virtuale Demenze), composed of sixteen IRCCSs. One of them was not involved in the present study because it deals with pre-clinical research. The IRCCS skills within a Virtual Institute are brought together to address, with harmonized diagnostic and therapeutic methodology, a range of pathologies that are relevant from an epidemiological point of view.

The task of extracting relevant information from eCRFs is commonly referred as Information Extraction (IE), also known as text-mining, and it has the goal of making explicit the semantic structure of a text, so that we can make use of it [5]. The simplest technology to deal with human-written text are regular expressions (regex) [6], i.e., sequences of characters that specify a text pattern, typically used to perform string-search operations. An example of a regex application can be seen in the Supplementary materials. Regexes, in spite of being extremely powerful for well-defined formats, may lack adaptability when it comes to expounding upon documents that are written in free text format. Complex documents require increasingly complicated algorithms, because getting information from a free text can be extremely challenging [42]: modern tools exploit Artificial Intelligence (AI) by means of statistical models [7] or, in the last decade, Deep Learning (DL), using multiple layers to progressively extract higher-level features from the input [8]. When applied to human-written texts, they are referred to as Natural Language Processing (NLP). One example is Question Answering (QA), which has the goal of finding answers to human-written questions. The advent of the Transformer architecture [9] allowed the NLP scientific community to create increasingly effective models. Some of the most famous architectures are BERT [10], T5 [11], and GPT [12]. These models are typically crafted with a two-step process: the first is pre-training, an unsupervised procedure where the model is fed a colossal amount of unlabeled text (e.g., BERT corpus is composed of 3.3 billion words); the second step is fine-tuning, a supervised training where the model is fed a relatively small amount of labeled training examples (e.g., a famous QA dataset, the Stanford Question Answering Dataset, SQuAD [13], is composed of a hundred thousand examples), and learns to perform a specific task. One of the main limits of this process is that it requires an enormous amount of text in the pre-training phase (tens/hundreds of billions of words, ideally), thus models available in literature are often trained on generic corpora, usually striving when it comes to specific topics. There have been efforts to overcome this limitation; Biomedical BERT (BioBERT, [14]), one of the most famous and successful ones, outperformed the original BERT on several biomedical NLP tasks. Another major

constraint is that the vast majority of available models are trained on English corpora. The so-called less-resourced languages, e.g., Italian, are underrepresented in this scenario. However, in the recent years efforts have been made to address this problem, bringing on the one hand to the development of multi-language models [15] [16], and of models specific for languages different from English [17] on the other. To complete the landscape of NLP technologies currently available, in the last year several Large Language Models (LLMs) have been publicly released in the NLP ecosystem. One famous example is ChatGPT, released at the end of 2022, a conversational AI system based on the GPT language models [18]. After that, other open-source LLMs have been released (e.g.: Vicuna, available at https://huggingface.co/lmsys/vicuna-13b-v1.5, and Falcon, available at https://huggingface.co/tiiuae/falcon-180B) as well. These models are usually trained on massive corpora (dozens of Terabytes of texts), and fine-tuned on very specific tasks, making them the state-of-the-art chatbots. The most peculiar difference with respect to previous models is that LLMs are generative, which means that they can elaborate information in the presented prompt or in the original training dataset and generate new information. However, LLMs typically present some critical issues: many of the current models (e.g., ChatGPT) can only be accessed through Application Programming Interfaces (APIs), forcing the user to send all information to private servers. This could have major ethical implications, considering the Health Insurance Portability and Accountability Act (HIPPA) and the General Data Protection Regulation (GDPR) [26]. In addition, they are trained on broad subject corpora, and could therefore struggle when it comes to very specific topics.

This work focuses on the clinical and technological decisions made to create a harmonized CRF within the IVD, along with developing a pipeline for extracting relevant information from eCRFs. The results include a common eCRF shared among fifteen independent hospitals all over Italy, and the implementation of an automated CRF-to-DB NLP pipeline. The primary aim of this work was not to devise new NLP models, but to distribute a software to enable efficient IE from eCRFs collected as part of a multi-center initiative. The ultimate goal is to fully utilize valuable information contained in hospitals' clinical documents, often unused due to labor-intensive manual retrieval process. In the following we will describe the methodological decisions that led to the implementation of the eCRF consensus and NLP Extraction and Management Tool (NEMT, the software implementing the end-to-end pipeline), the experiments carried out to test its performance, and we will discuss NEMT results and its limitations, while giving insights for possible future implementations.

II. Methods

The overall process, from patient to structured data, was divided into three tasks:

- IVD eCRF consensus: definition and implementation;
- Information extraction: semi-automated data extraction from the clinical document;
- IVD Database: automated data conversion from unstructured to structured.

IVD eCRF Consensus

The first step of the process has been the crafting of the CRF consensus format. This activity was performed on two types of CRF: clinical (i.e., comprising mostly medical information) and neuropsychological (i.e., containing scores and a report on the patient's cognitive performance in standardized neuropsychological tests). It is important to note that the IRCCSs of the IVD are located in seven different Italian regions, each certifying and accrediting different eCRFs providers, thus it was not possible to use a single electronic medical chart for the IVD. However, the conceptual application of the CRF consensus was implemented in fifteen hospitals, so that all the requested patients' information was obtained and recorded in similar but different sections and ways for each provider. The development of the proposed pipeline had to take into account this fact. Regarding the clinical eCRF, a first draft was produced by the "Clinics" task leader of the IVD (P.T. from IRCCS Istituto Neurologico Carlo Besta). The draft was then evaluated by each delegate from the other IVD Institutes involved in the task (one person per Institute). Their comments and suggestions were then discussed in further meetings. The result of this process was the creation of a final consensus CRF, to which every Institute had to uniform its assessments and records. Whenever possible, the consensus CRF was transferred directly to the Institutes' electronic medical charts. In some Institutes, however, this was not possible, either because there was no electronic chart for inpatients yet or because the existing electronic chart did not allow for a major structural change. In these cases, the CRF was conceptually implemented so that all information was collected and recorded in different sections in the clinical charts. The creation of the neuropsychological consensus CRF followed a similar process. The first draft was prepared by C.M. from IRCCS Policlinico A. Gemelli Foundation, the leader of the neuropsychology working group. The complete list of items of the two eCRF consensuses can be found at the following link: https://doi.org/10.5281/zenodo.11104006

Information Extraction

NEMT

The goal of this task was to develop a tool to automatically extract CRF Consensus items from the IVD centers' eCRFs. The main difficulty was the lack of a common structured template. After IE, these items had to be converted into an organized structure and stored in a DB. For this process, we developed NEMT, whose block diagram is shown in Fig. 1. NEMT is a web-based software that adheres to the principles of "privacy by design and by default". It has been deployed locally in each center of the IVD on virtual machines (VMs) implemented with VMware technology. The VMs run Ubuntu 20.04, with 8 GB RAM, 8 virtual cores, and 50 GB hard disk. The only input required for NEMT is the clinical/neuropsychological eCRF in PDF format. The complete workflow, from patient assessment to upload data to REDCap, is described below.

First, the patient is visited by the clinician, who creates an eCRF using the hospital's electronic medical record. The informed consent is obtained during the assessment (IRB Approval Numbers: 74/2020 and 10/2022). The clinical document is preprocessed by NEMT: the text is extracted from the PDF by removing headers, footers, and unnecessary parts of the document. The items of the consensus are then extracted using the text-mining algorithm, which combines regexes for well-defined data and NLP for open questions. It is important to mention that all operations cited so far are performed on the hospital's local computers, without sending any data to the outside world. Finally, the data are converted into a well-defined structure and sent to the remote REDCap using PyCap (https://github.com/redcap-tools/PyCap), a Python module that exposes the REDCap API. In order to maintain "privacy", the patient's fiscal code is encrypted before being sent. In addition, the patient's identity is stored locally in a "transcoding table" to enable subsequent re-identification, if required. We opted for a locally-distributed solution, instead of a single remote server, so that the eCRF, which contain sensitive patients' information, are elaborated on the local machines of every center. This solution avoids the transmission of patient's information via Internet and possible data leaks.

The core of NEMT is the IE algorithm, represented by Block 2. It combines a regex and NLP approach, making it flexible enough to handle eCRFs crafted by different centers. For the NLP model we decided to use a QA architecture, because its end-to-end nature allows direct text extraction from eCRFs without further elaborations. The QA approach was used only for clinical CRFs, because neuropsychological ones are almost entirely tabular, making the regex approach suitable for reaching a high level of accuracy. The QA model is explained in the Question Answering approach section. The IE task is semi-automated: extracted items are presented to the user for confirmation (i.e., human in the loop), so that mistakes can be fixed before sending data to the DB. This step is necessary, because NLP models can produce incorrect extractions [41], thus a human quality check is always advised. Indeed, some human interventions is still required in the IE process, but it is limited to checking the quality of the algorithm output, while the burdensome and tedious extraction task is performed by NEMT. With this paradigm, AI is a sort of "assistant", while the final decision is always in the hands of the clinicians. As highlighted by several recent studies, AI algorithms are

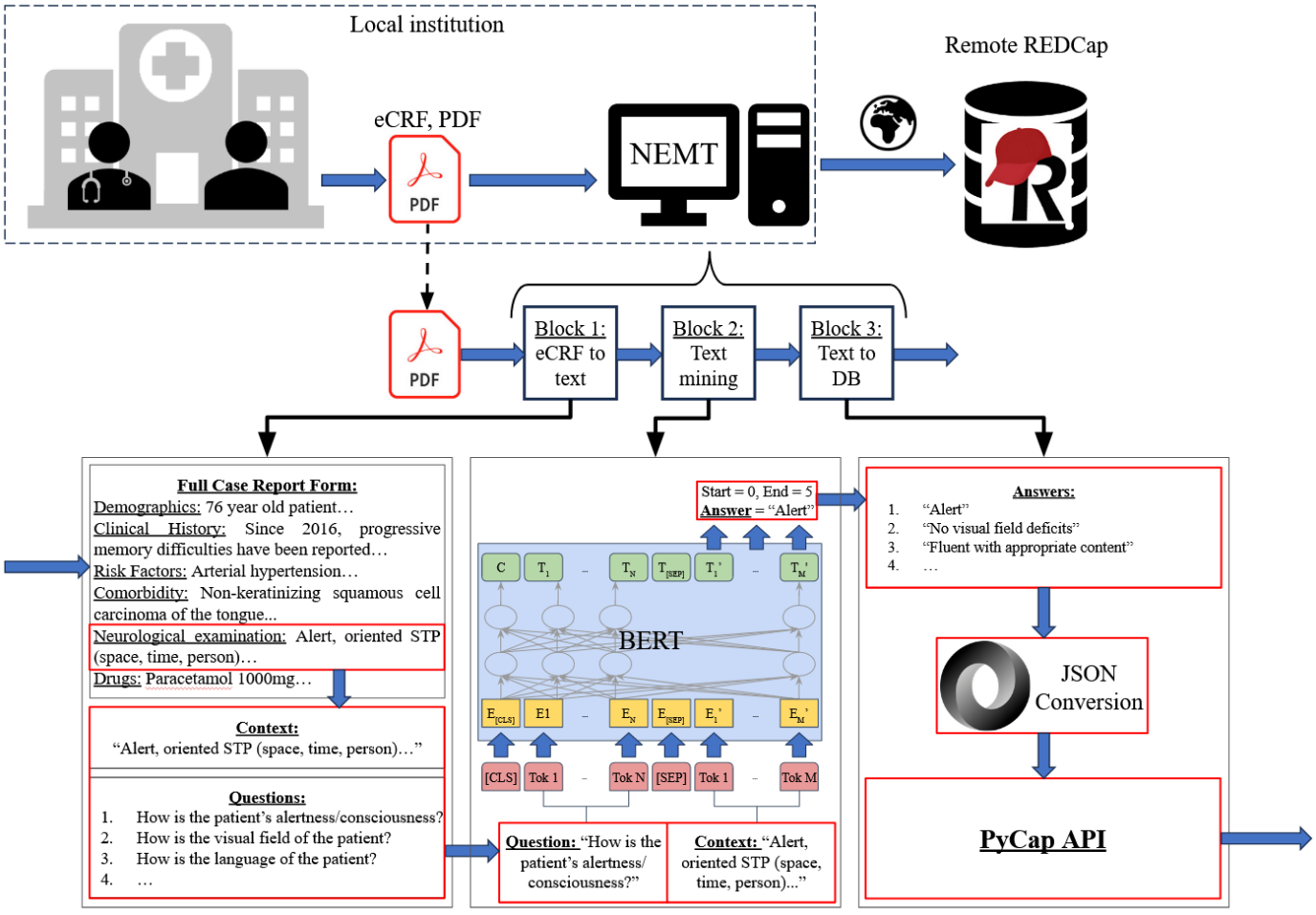

Fig. 1. Complete workflow, from eCRF to REDCap storage. The upper part of the figure shows the high-level actions performed by the clinician: first, the clinician visits the patient, then they use NEMT by inputting the eCRF in their local version of the software; the extracted data are then automatically sent to the remote REDCap database. The lower part of the figure shows the main blocks of NEMT. Block 1 converts PDF eCRF to text and extract the content related to the Neurological examination. Block 2 applies the QABot text-mining pipeline to answer the questions; for every question, the answer (if possible) is the span of the context that represents the answer to the question; if impossible, there is no answer. Block 3 organizes data in a JSON structure and sends them to the remote REDCap by means of the PyCAp API.

becoming part of the clinical decision-making process, speeding up the whole tasks and contributing to lower the cost of medicine [39] [40] through significant optimizations. It is important to note that all technologies used to implement NEMT are either open-source (e.g., BERT models) or free (e.g., REDCap). This is a key factor for IRCCSs, as it allows to implement effective solutions while promoting accessibility.

Finally, it is important to consider that the ultimate goal of NEMT is to expedite the IE process so that clinicians can avoid a long, and repetitive task, resulting in time savings. To quantify it, we conducted a test comparing the time to extract 25 clinical and 25 neuropsychological eCRFs from 5 centers by humans and NEMT (also taking into account the time for corrections). The results are described in the NEMT paragraph of the Results section.

Question Answering approach

Considering the extremely rapid pace at which the situation is evolving, we decided to found our pipeline on a well-established architecture, i.e., QA BERT-based models, also known as QABots.

The input of a QABot is the concatenation of a human-written question and the context, a human-written text containing the information that the model will elaborate to retrieve the required data. Question and context are concatenated into a single sequence, which is then tokenized, using WordPiece tokenization (with a vocabulary of approximately 31 000 tokens), adding special tokens [CLS] at the beginning and [SEP] to separate the question and the context. After that, there is the embedding layer, which converts tokens by mapping them into 768-dimensional dense vectors, forming the basis for subsequent processing, i.e., position embeddings, that captures positional information up to 512 tokens, and token type embeddings, which is crucial for tasks involving sentence pairs. Layer normalization and dropout stabilize and regularize the embedding outputs, enhancing model robustness. At the core of the model is a 12-layer Transformer encoder, each one composed of sub-layers designed to perform self-attention and intermediate transformations, followed by a dropout layer to prevent overfitting. The intermediate layer, comprising a dense layer that projects data from 768 to 3072 dimensions and a Gaussian Error Linear Unit (GELU) activation function, enhances non-linearity. The output layer involves a projection back to 768 dimensions, coupled with normalization and dropout, ensuring consistency in data dimensions and regularization. The QA head, added on the top of the encoder during the fine-tuning, processes the encoder outputs to predict the start and end positions of the answer span, making the model able to handle question answering tasks.

The starting checkpoints for our model are the ones developed in [19] and [20]:

- BioBIT (https://huggingface.co/IVN-RIN/bioBIT), a BERT-based checkpoint pre-trained on the Italian translated version of the original BioBERT corpus;
- MedBIT (https://huggingface.co/IVN-RIN/medBIT) and MedBIT-r3-plus (https://huggingface.co/IVN-RIN/medBIT-r3-plus), created starting from BioBIT and pre-trained once more with small, high-quality natively Italian biomedical corpora.

These models can be downloaded and exploited locally, avoiding the submission of data to external cloud-based ecosystems and related privacy problems. Moreover, these checkpoints can be fine-tuned on GPU-based machines. To fine-tune a QA model, usually a few thousand examples are required. A QA example is composed of three parts: a context, a question related to the context, and an answer, i.e., the span of the context that answers the question. The original SQuAD1.0 dataset, which constitutes the standard reference when dealing with QA datasets, is composed of approximately a hundred thousand examples. It is fundamental to notice that these models are extractive, meaning they cannot generate new information by inferring data, but they are limited to extracting it from the original context. The 2.0 version of SQuAD [22] adds over 50 thousand unanswerable questions, enabling models to detect impossible questions, while a SQuAD1.0-based one will always predict an answer.

Inter-Annotator Agreement

Inter-Annotator Agreement (IAA) measures the level of agreement among annotators when answering questions. Evaluating IAA is a crucial step in ensuring the quality of labeled datasets. Determining which IAA values are considered high is complicated and depends on many factors. Generally, a value below 20% is considered poor, 20-40% is fair, 40-60% is moderate, while values above 60% are good, and above 80% excellent. IAA can be calculated in several ways; we used Exact Match (EM) and F1-scores, in order to have values comparable to some of the fine-tuning process performance. EM-score ranges from 0 to 1, indicating the ratio of perfectly matching answers between the gold standard and annotator. F1-score is the harmonic mean of Precision and Recall, which are measured by counting shared tokens (typically a subpart of a word) between answers and the gold standard, thus obtaining True Positives (TPs), False Positives (FPs), and False Negatives (FNs). To calculate the F1-score on several questions, we used Macro-F1, obtained by averaging the single F1-scores.

Real-world IAA values vary widely across different datasets. For example, Yadav et al. [30] present a modest value of 58% while annotating semantic relations between two elements of noun-noun compounds, while Dinh et al. [31] reach much higher values of 91-94% while annotating antibody and antigens on their corpus. Roberts et al. [32] results reflect this significant variability while annotating questions about patients in electronic health records, with values ranging from 61% to 88%. This variability suggests that IAA depends on problem complexity, and even highly qualified clinical staff may yield relatively low values.

To calculate IAA, we randomly selected five of the fifteen centers of the IVD. Each of these five centers identified one clinician who worked on two eCRFs for each institute. Since the annotation process is time-consuming, we have estimated that ten documents represent a favorable ratio between the effort and the amount of generated data. For every document, the gold standard are the annotations of the clinician of the center that produced the document. For the complete results see

Supplementary materials at section "IAA full results".

Dataset creation and training

We created a QA dataset by collecting data from eCRFs of each IRCCS of the IVD. Every selected center annotated sixty clinical and sixty neuropsychological documents containing items of the CRF Consensus. Starting from these annotations, we created two datasets:

- Clinical IVD dataset, composed of about 60k QA examples in total;
- Neuropsychological IVD dataset, composed of about 36k QA examples in total.

Data may be made available via request. Data sharing will require a formal data sharing agreement and approval from ethics committees involved.

We exploited these datasets to train the BERT-based models mentioned above. In order to increase the size of the datasets, and possibly enhance the model performance, we combined IVD datasets with two datasets. The first was Biological Question Answering (BioASQ) [21]. BioASQ is a research initiative, organized as a series of challenges in the field of biomedical semantic indexing and QA. This dataset is made up of datasets from several challenges (4^{th}, 5^{th}, 6^{th} version), and comprises a total of around 8 thousand examples. The second dataset was SQuAD. Indeed, not all Consensus CRF questions required in-depth medical experience to be answered; some of them can be handled by understanding the grammatical structure of the context. For these reasons, we included the SQuAD dataset. Although not biomedical, it contains examples of generic questions, potentially improving the efficiency of the models. The original datasets were in English, so we translated them using the Google's neural machine translation system [29]. The quality of the translation has been rigorously evaluated by experts in the field (C.C., A.R.), Italian mother tongue and high proficiency English speakers.

We ran 3 fine-tuning experiments:

1. We fine-tuned Bio/MedBIT on the IVD dataset;
2. We fine-tuned Bio/MedBIT on a dataset created by merging the IVD dataset with BioASQ SQuAD;
3. We fine-tuned Bio/MedBIT on a dataset created by merging the IVD dataset plus BioASQ SQuAD and generic SQuAD 2.0.

The fine-tuning was performed using the Deepset Cloud libraries (https://docs.cloud.deepset.ai/) at the IRCCS Centro San Giovanni di Dio Fatebenefratelli high-performance computing center, which was equipped with four A100 GPUs, each of them with 40GB of RAM, and took a total of approximately 96 hours. The fine-tuning parameters were: a batch size of 10, learning rate of 1e-5, 3 epochs, warmup ratio of 0.2, and maximum sequence length (i.e. the maximum sequence token length of one input text for the model) of 512.

The fine-tuned models were evaluated using the EM and F1 scores previously defined in the Inter-Annotator Agreement section, and two other metrics: Lenient Accuracy (LAcc) and Mean Reciprocal Rank (MRR), commonly used in QA systems [21], [33]-[35]. Strict Accuracy (SAcc), also a common metric used to evaluate QA systems, has not been calculated, as it is very similar to EM score. LAcc scores a question as correct if the gold standard (defined in Section Methods - Inter-Annotator Agreement) is included among the first "k" candidate answers, where "k" was set to five. The RR of a single question is the reciprocal position of the correct answer among the k candidates (e.g., if the correct answer position is 3, its RR is 1/3), while the Mean RR is the average of the RR on all answers. As additional information, we added the representation of feature spaces at different layers of the model, analyzed with a Manifold Discovery and Analysis (MDA) tool [43], to show the model ability to learn proper features during the fine-tuning process. This can be found in Supplementary materials in the section "Networks feature space".

Large Language Models

Given the concerns surrounding the privacy of sensitive patients' data, we decided not to implement LLMs in our pipeline. However, for research purposes, their performance was evaluated and compared with that of BERT-based fine-tuned models, on the Test split of the fine-tuning dataset. In particular, we tested two LLMs:

- ChatGPT, a state-of-the-art generative LLM, based on the GPT architecture. Built on a Transformer network with hundreds of billions of parameters, ChatGPT is trained on diverse and extensive corpora, making it able to comprehend and generate human-like text across a wide range of domains. For this work, the free version of ChatGPT was used to perform test on 10manually anonymized eCRFs;
- Vicuna [23], an open-source chatbot trained by fine-tuning LLaMA (a 7-65B parameters LLM trained exclusively on public datasets [24]) on user-shared conversations collected from ShareGPT [25]. Preliminary test shows Vicuna-13B achieves more than 90% quality of ChatGPT. Although it seems promising, authors state that further tests are required.

IVD Database

The last step of the process implies the organization of unstructured data into a relational DB. The platform chosen to host the DB is Research Electronic Data Capture (REDCap, official website https://www.project-redcap.org/) [27] [28], a secure, web-based solution designed to support data capture for research studies. REDCap provides advanced security measures that ensure data confidentiality in compliance with regulatory standards, e.g., GDPR. REDCap provides APIs that allow interaction with external software, making it possible to develop pipelines that automatically store information in the DB without the user having to impute them manually. This fact greatly enhanced the speed of the overall process within the IVD, making the data storage operation far less burdensome, and thus more acceptable for clinicians.

III. Results

eCRF Consensus of IVD

The result of this work was the definition of two lists of items, one clinical and one neuropsychological. These lists represent the minimum dataset that each eCRF had to provide, regardless of the institute in which the patient was assessed. The clinical eCRF consists of six sections, namely:

1. Demographics: sociodemographic details about the patient (e.g., date of birth, sex, education, occupation);
2. Clinical history: cognitive, psychopathological, and motor symptoms of the patient (e.g., symptoms reported by the patient and the informant);
3. Risk factors: an in-depth list of dementia-related pathologies across the patient's family, and other risk factors (e.g., physical activity, smoking);
4. Comorbidity: list of current/past comorbidities (e.g., hypertension), and history of psychiatric disorders;
5. Neurological examination: objective biological parameters of the patient (e.g., weight, blood pressure), and a list of neurological examination findings (e.g., patient's state of consciousness);
6. Drugs: the list of medication the patient was taking at the time of the examination and prescribed medication for treatment after discharge.

The neuropsychological eCRF consists of three sections:

1. Demographics: sociodemographic information about the patient (e.g., date of birth, sex, education);
2. Behavioral and functional assessment: results of exams that assess the ability of an individual to perform basic self-care tasks, and the severity of neuropsychiatric symptoms commonly observed in dementia patients (e.g., agitation, aggression);
3. Neuropsychological assessment: results of exams that assess cognitive impairments of the patient (e.g., orientation, memory, visuo-spatial skills).

Information Extraction

NEMT

The results show that for clinical eCRFs the average human IE process takes 1080 ± 360 seconds, whereas with NEMT it is reduced to 390 ± 120 seconds. For neuropsychological eCRFs, humans require 210 ± 180 seconds compared to 30 ± 21 seconds with NEMT.

Inter-Annotator Agreement

The QA approach was only used for clinical CRFs, as the regex approach was sufficient to achieve an accuracy level of over 85% on mostly tabular neuropsychological CRFs (the full results can be found in the Supplementary materials in the section "Neuropsychological eCRF Information Extraction"). For this reason, the IAA was not calculated for the neuropsychological eCRFs.

IAA was measured by means of two metrics: EM-score and macro F1-score. The "EM score" column has three sub-columns: "No answer", with scores for impossible questions (thus with an empty answer), "Text answer", with scores for questions with non-empty answers, and "Overall", the average for all questions. At first, they were calculated at the document level, as shown in TABLE I. A second experiment was conducted by dividing the CRFs into sections. The idea is that, in case of very different results from different sections, we could apply the QABot only on specific parts of the CRF, and not on the whole document. Results are reported in TABLE II. IAA has very different values in different sections. "Clinical history" and "Risk factors" have a fair value, "Comorbidity" is moderate, while "Neurological examination" is good and "Drugs" section has an excellent IAA.

Dataset creation and training

Based on the results of the IAA, we decided that the NEMT QABot model would only work on items of the "Neurological examination" section of the eCRF. This is due to the fact that the IAA F1-scores for the first three sections (Risk factors, Clinical history, and Comorbidity) did not reach the "good IAA" threshold of 60%, while the Drugs section is often structured on IVD CRFs, making the implementation of an NLP tool unnecessary. For this reason, the annotated documents produced by the IVD centers were divided into sections, and only Objective exam was used for the actual fine-tuning. The number of items corresponding to this section was 25, resulting in a total of about 20 500 examples for fine-tuning. These examples have been randomly split into training and test set, corresponding to 90% and 10% of the total, respectively. The training set was further split into evaluation and train set, corresponding to 20% and 80% of the total training set. To summarize, the train set counted about 15 000 examples, the evaluation set 3500, and the test set 2000. Then, BioASQ and SQuAD examples have been added to the IVD dataset (train, evaluation, and test set), in order to increase the general applicability of the models. It is important to note that, while examples for the IVD dataset were taken from the Neurological examination section of the CRFs, with the list of 25 questions, both for BioASQ and SQuAD dataset they typically consisted of a relatively short context (1 or 2 sentences) and a single question. Data are summarized in TABLE III. The total number of examples is:

- IVD dataset: 20 544 examples → Train set ~ 15 000, Evaluation set ~ 3500, Test set ~ 2000
- IVD + BioASQ dataset: 28 458 examples → Train set ~ 20 500, Evaluation set ~ 5000, Test set ~ 3000
- IVD + BioASQ + SQuAD dataset: 90 622 examples → Train set ~ 65 000, Evaluation set ~ 16 000, Test set ~ 9500k

The fine-tuned model with the highest performance, called bioBIT_QA from now on, has been shared on the HuggingFace hub at the following link: https://huggingface.co/IVN-RIN/bioBIT_QA. The model can be downloaded for free and used with the Deepset libraries.

Error analysis

To better understand the performance of bioBIT_QA, we examined the type of errors it made. A random 10% of the IVD test set, corresponding to 9 eCRFs (225 questions), was processed. We defined four error types, reporting subsequently for each: the context, the question, the reference answer (underlined) and the answer predicted by the bioBIT_QA (in bold):

1. Not pertinent answer (2.70% of the total): the model gives an answer unrelated to the gold standard, e.g.:

Context: "*The patient is awake, cooperative, partially oriented in space (knows the city but **confuses the name** of the hospital) and temporally oriented (does not remember the day of the week).*"; Question: "*How is the patient's alertness/consciousness?*"; Reference answer: "*awake*"; Predicted answer: "***confuses the name***".

2. Not impossible answer (70.27% of the total): the QABot marks the question as impossible, but the gold standard contains answer, e.g.:

Context: "*Cranial nerves: normal vision, no visual field deficits (quadrant comparison test), normal eye movements, remaining cranial nerves normal.*"; Question: "*How is the visual field of the patient?*"; Reference answer: "*no visual field deficits (quadrant comparison test)*"; Predicted answer: impossible.

3. Short answer (18.92% of the total): the model identifies a shorter answer than the gold standard, thus missing part of the reference answer, e.g.:

Context: *"**Alert**, oriented STP (space, time, person). Speech is fluent and appropriate in content. Mood aligned.*"; Question: "*How is the patient's alertness/consciousness?*"; Reference answer: "***Alert**, oriented STP*"; Predicted answer: "***Alert***".

4. Long answer (8.11% of the total): the model identifies a longer answer than the gold standard, thus including text not related to the reference answer, e.g.:

Context: *"Alert, oriented STP (space, time, person).* ***Speech is fluent with appropriate content****. Mood aligned.*"; Question: "*How is the language of the patient?*"; Reference answer: "*fluent with appropriate content*"; Predicted answer: "***The speech is fluent with appropriate content***".

The vast majority of errors are of type 2 (approximately 70% of the total), which means that bioBIT_QA identified the question as impossible, even though the reference actually contained an answer. This could have several causes, namely:

- The gold standard answer contains an acronym and the model misses its meaning. We report an example where the annotator marked the answer as the acronym "EOM" (ExtraOcular Movement) and bioBIT_QA missed it, e.g.: Context: "*Head and globes aligned, IOM and EOM within normal limits, intact cranial nerves in comparison.*"; Question: "*How are the patient's extraocular movements?*"; Reference answer: "*EOM within normal limits*".
- The gold standard answer is vague and not directly related to the question. We provide a case where the annotator identifies a detail regarding the patient's ability to move, but which does not directly address the question, e.g.: Context: "*Bradykinesia on tapping and adiadochokinesia. Need for assistance to rise from chair. Walking with supportfrompersonnel, with shortshufflingsteps. Freezing of gait.*"; Question: "*What is the patient's posture?*"; Reference answer: "*Need for assistance to rise from the chair*".
- The gold standard answer contains typos or misspellings. We report the same example as before, in which the annotator gave as an answer a piece of text that contained missing spaces between several words due to typos. It is interesting to note that the model predicted the correct answer once the text has been corrected, indicating that the errors can sometimes be caused by the writing errors, e.g.: Context: (see previous example); Question: "*How is the patient's gait?*"; Reference answer: "*Walking with support from personnel, with short shuffling steps*".

Overall, these discrepancies highlight a lack of consistency in the annotation tasks. Paying more attention to this fundamental process could lead to a better-quality dataset and therefore higher overall IE performance.

Large Language Models

For research purposes, two LLMs, namely ChatGPT v3.5 (3 August 2023 version) and Vicuna, were tested on the same task of the fine-tuned QABots. We chose version 3.5 of ChatGPT because the online version is free to use. The tests were performed with 10% of the three test sets, and then the EM and F1-scores were calculated and compared (results are shown in TABLE IV). It was not possible to calculate LAcc and MRR, because LLMs give a single response when prompted with a QA style. It is fair to point out that comparing a generative model to the same metrics used for an extractive model could be misleading, as the addition of words typical of LLMs, even if meaningful, could lead to a drop in performance. For this reason, we performed several tests to find the best prompt so that the model would not change the original context. The prompt used for this test was the following:

Starting from this text:
"CONTEXT"
Answer to the following questions, without editing the original context. If an answer is not present, write the answer "Not present in the original context". Format the answers as a JSON:
"LIST OF QUESTIONS"

TABLE I
IAA ON THE COMPLETE DOCUMENTS USED FOR TEST, CALCULATED BY CENTER, AND OVERALL

	EM score [%]			F1-score [%]
	No answer	**Text answer**	**Overall**	**Macro**
Center 1	78.1 ± 9.4	29.2 ± 14.8	51.9 ± 15.9	72.1 ± 16.4
Center 2	68.2 ± 15.5	40.9 ± 13.0	54.1 ± 10.9	80.7 ± 9.1
Center 3	87.3 ± 8.9	33.8 ± 14.4	59.3 ± 10.0	78.3 ± 8.5
Center 4	78.1 ± 9.7	39.9 ± 8.9	57.2 ± 6.1	80.9 ± 5.2
Center 5	87.4 ± 7.2	34.9 ± 11.1	57.6 ± 7.1	81.6 ± 9.1
Overall	**79.9 ± 11.8**	**35.7 ± 12.0**	**56.0 ± 10.0**	**78.7 ± 9.7**

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2024.3450118

TABLE II
IAA ON THE DOCUMENTS USED FOR TEST, SPLIT INTO SECTIONS

	EM score [%]			F1-score [%]
Section	**No answer**	**Text answer**	**Overall**	**Macro**
Clinical history	38.8 ± 24.1	12.7 ± 13.5	25.7 ± 23.3	38.1 ± 22.9
Risk factors	74.6 ± 20.4	0.8 ± 1.3	37.7 ± 40.9	34.0 ± 3.4
Comorbidity	23.3 ± 38.3	10.0 ± 5.0	16.7 ± 25.8	51.1 ± 4.6
Neurological examination	22.6 ± 20.2	28.9 ± 15.5	25.7 ± 18.1	64.8 ± 19.2
Drugs	28.3 ± 24.7	35.8 ± 21.8	32.1 ± 21.2	85.9 ± 9.9

TABLE III
TRAINING SCORES FOR THE THREE STARTING CHECKPOINTS (BIOBIT, MEDBIT, MEDBIT-R3) ON THREE DIFFERENT DATASETS (IVD, IVD + BIOASQ, IVD + BIOASQ + SQUAD)

Starting model	**Fine-tuning dataset**	**EM score [%]**			**F1-score [%]**	**LAcc**	**MRR**
		No ans	**Text ans**	**Overall**	**Macro**		
BioBIT	IVD	97.5	52.2	73.7	81.8	0.825	0.771
	IVD + BioASQ	97.7	66.5	**78.1**	**84.7**	**0.834**	**0.810**
	IVD + BioASQ + SQuAD	97.2	63.3	75.7	82.1	0.671	0.616
MedBIT	IVD	96.9	50.8	72.7	81.0	0.820	0.758
	IVD + BioASQ	97.8	64.4	76.9	82.8	0.785	0.751
	IVD + BioASQ + SQuAD	97.9	63.9	76.4	82.8	0.654	0.588
MedBIT-r3	IVD	98.6	82.4	73.5	80.3	0.824	0.767
	IVD + BioASQ	96.9	65.8	77.4	84.3	0.820	0.794
	IVD + BioASQ + SQuAD	98.2	63.5	76.3	82.7	0.682	0.622

TABLE IV
RESULTS OF IE EXECUTED BY MEANS OF LLMS CHATGPT AND VICUNA

Starting model	**Test set**	**EM score [%]**			**F1-score [%]**
		No answer	**Text answer**	**Overall**	**Macro**
ChatGPT (v3.5)	IVD	76.6	29.4	51.0	33.3
	IVD + BioASQ	84.3	58.9	**69.9**	**52.5**
	IVD + BioASQ + SQuAD	87.2	48.9	61.2	49.3
Vicuna (v13b)	IVD	74.6	23.8	44.1	36.9
	IVD + BioASQ	88.9	48.9	**66.3**	**49.5**
	IVD + BioASQ + SQuAD	86.2	38.9	54.1	39.5

IVD Database

At the time of writing, NEMT was being used in the IVD to process and store items from two hundred eCRFs. These data could be used to test new scientific hypotheses and understand the onset mechanisms of Alzheimer's Disease (AD) and other forms of dementia. The data could also be used in future clinical research studies.

IV. DISCUSSION

IVD eCRF Consensus

From a clinical point of view, the most important achievement of the clinical and neuropsychological eCRF is the standardization and systematic collection of clinical data collection and recording among the 15 participating Institutes, which are spread across the entire national territory. This means that patients with cognitive impairment (and, in prospect, those with other neurological diseases) will be neurologically and neuropsychologically evaluated in the same manner across the country, which is important both from a healthcare perspective and for research studies requiring large cohorts with homogeneous and well-curated datasets. In addition, the implementation of standard eCRFs will increase the completeness and granularity of clinical and neuropsychological assessments in contexts where less detailed examinations were previously performed.

Information Extraction

NEMT

The results show that IE is about three times faster in clinical eCRFs, while it is seven times faster in neuropsychological eCRFs compared to humans. Although the overall performance of NEMT could still be improved in the future, these results show that it leads to significant time savings for medical staff.

Inter-Annotator Agreement

The overall IAA EM score is 56%. In addition, the overall EM-score for empty answers is close to 80%, i.e. 4 out of 5 times annotators agree that a question cannot be answered. The

macro F1-score of 79% is a good result; it means that, on average, ~80% of the annotators' answers overlap with the gold standard. Given the complexity of the contexts and questions, the results are in line with the literature. Nevertheless, they show the complexity of the problem, with some very specific questions that correspond to convoluted, highly variable, grammatical structures in the answers.

Regarding the IAAs of the individual CRF sections, three of them (i.e.: Clinical history, Risk factors, and Comorbidity) have a very low EM-score for non-empty answers, ranging from 0.8 to 12.7%. This proves that even for human annotators it is difficult to give a reliable answer. F1-scores of these sections are relatively low as well, ranging from 34.0% (low) to 51.1% (moderate). While EM-scores are relatively low even for the remaining two sections (Neurological examination and Drugs, with 28.9-35.8% for non-empty answers and 25.7-32.1% overall) their F1-scores are relatively high (64.8% and 85.9%), proving that annotators reached a good/excellent agreement. This could be due to the objectivity of these sections: while items like risk factors and comorbidity are less defined and more open to interpretability of the single rater, results of objective exams and drugs are objectively defined and thus easier to annotate in a correct way.

Dataset creation and training

The best results are provided by the BioBIT model fine-tuned with the IVD and BioASQ datasets merged together: EM = 78.1%, F1 = 84.7%, LAcc = 0.834, MRR = 0.810; this pooled dataset provides the best scores for the other two starting checkpoints as well. This is likely because all the questions are biomedical, so IVD examples could increase the efficacy of BioASQ questions, and vice versa. Conversely, the results obtained by adding the SQuAD dataset have a slightly lower score (worst EM score difference = -2.7%, worst F1-score difference = -2.6%, worst LAcc = -0.163, worst MRR = -0.194). This proves that the addition of a generic QA dataset in the NEMT training pipeline does not improve the performance of the fine-tuned models, in this specific case. Thus, data quality is preferable over quantity [19]. It is interesting also to note that both the EM and F1-scores are higher than the corresponding IAA values: the EM score is 78.1% versus 25.7%, and the F1-score is 84.7% versus 74.6%. This may come as a surprise, as it is generally assumed that the IAA represents the upper limit of the performance that the model can achieve: if humans cannot agree on the labeling, we can assume that NEMT does not do better. However, there are studies indicating that this is only an untested assumption and there is no authoritative source to support it [36]. While a high IAA is desirable for creating reliable training datasets, it does not directly determine the upper limit of performance that an NLP system can achieve. Performance can vary significantly depending on the complexity of the task, the quality and size of the training data, the model architecture, and other factors. The performance of LAcc and MRR follows a similar pattern when using IVD and IVD + BioASQ datasets.

These results demonstrate the effectiveness of the proposed data collection and fine-tuning strategy, with our best performance being higher than in other developed QA systems [33]-[35]. However, it is important to remember that the model we propose is thematically very narrow, while the other works mentioned have developed biomedical tools for general purpose. By incorporating bioBIT_QA, NEMT was able to extract information from eCRFs from fifteen different institutes with good performance. These data were processed and efficiently stored in the REDCap DB, so that this structured version of the IVD data can be used in the future.

Large Language Models

The results of the assorted datasets are quite different. Due to the single answer given by the LLMs when prompted with a QA style, it was not possible to calculate LAcc and MRR. The dataset with the lowest results, for both ChatGPT and Vicuna, is the IVD dataset. It is important to remember that this dataset contained the longest contexts (an entire paragraph of a medical report), ranging up to several hundred words. For this reason, it is probably the most difficult dataset to work with for general-purpose models. In addition, it should be noted that generative capacity can manifest itself with both positive and negative consequences, since in some cases LLMs could generate information that is not present in the original text (the so-called hallucination phenomenon [37]), even if this was explicitly forbidden in the prompt. This was evident in the computation of the EM score (29.4% for ChatGPT and 23.8% for Vicuna. When extending the dataset with BioASQ and SQuAD, the performance increased, probably because the new examples had a much shorter context on average. ChatGPT achieved the best results on the IVD + BioASQ dataset with an EM and F1-score of 69.9% and 52.5%, respectively. The best results for the same dataset, which were also the best overall results, were obtained when we fine-tuned the NEMT BioBIT checkpoint. This achieved 78.1% for EM and 84.7% for F1-score, an increase of +8.2% and +32.2%, respectively, over ChatGPT. It is important to remember that ChatGPT and Vicuna are general purpose LLMs, and these results were obtained using a zero-shot approach [38]. Moreover, the adopted metrics are limited in some ways, as they compare the tokens of the reference and the predicted answer and ignore their semantic content. Nevertheless, the results show that for very specific contexts, BERT models pre-trained on corpora of relevant domains can perform better with appropriate fine-tuning, regardless of a much smaller number of parameters.

Best practices

Based on our work, we recommend the following best practices for IE from electronic health records in a multicenter initiative such as the IVD. First, defining and implementing a robust clinical consensus template for multicenter trials is essential to ensure minimization of missing data and efficient streamlining of data extraction. It is recommended to take as comprehensive medical history as possible, together with the cognitive history and social information reported from both the patient's and the caregiver perspective. Medical examinations must be thorough. We emphasize that clinicians should collect all necessary information regarding possible diagnostic suspicions. The family history and other relevant information

as well as lifestyle habits should be documented.

As far as the practical aspects of the pipeline are concerned, coordinating fifteen data centers spread across Italy, or possibly other countries, was not trivial. The simplest approach would be to centralize the processing and offer the IE pipeline as a service. This would also allow the use of more powerful hardware. However, this was not possible under the IVD initiative, as this approach would mean transmitting eCRFs over the internet, which would have required efficient anonymization of the data to prevent the potential disclosure of sensitive information. Therefore, we have pursued the adoption of the decentralized VMs developed in the hospitals' facilities. In addition, the implementation of an automatic software module for updating our QABot, in conjunction with an automatic log report, has greatly simplified the management of the NEMT software. We also recommend the use of OpenID connection protocol and the single sign-on system for user authentication and authorization. As a DB, we suggest the use of relational databases (e.g.: REDCap, or others), which are stable and well-suited for storing huge amount of clinical data.

V. Limitations and Future Work

The present work has limitations. First, the created QABot models were fine-tuned to a narrow topic specific for dementia, as highlighted also by the MDA analysis (see Supplementary material). Thus, applying them to other medical topics would require a new fine-tuning. Then, even though the training results were relatively high, the IAA was moderate. Although a high IAA does not directly imply good performance, it is desirable for the creation of reliable training datasets. In a future work, we could organize a training session for annotators, evaluate their agreement, and iteratively re-train them until a target IAA is reached. Furthermore, the current performance definition for both the IAA and the model scores takes into account the syntactic structure of the answers and not their semantic content. More clearly, assuming a context: "The patient shows symptoms of AD. We will subject him to a series of tests to prove the correctness of this diagnostic hypothesis (AD)" and the following question: "What is the diagnostic hypothesis for the patient?", if a reference annotator labels the text "Alzheimer's disease" as the answer and a second annotator marks "AD", then the IAA of this specific question would be zero for both EM and F1-scores. The same applies to a QABot that gives "AD" as the answer. While technically it is correct to assign a score of zero to this question, the different annotated answers have the same semantic content, since AD is the acronym for Alzheimer's disease, and thus their meaning is the same. One could therefore argue that the second answer is correct, from an IE perspective, because it does not matter if the DB contains in the "Diagnostic hypothesis" column the string "Alzheimer's disease" or "AD". This highlights the limitations of the metrics traditionally used to evaluate QABots and other NLP models in general. An interesting further development could be the definition of a new metric that takes this aspect into account and gives more importance to the semantic content of the predicted answers. Another limitation is represented by the adopted BERT-based architecture. Since the LLM breakthrough, benchmarks on every NLP task showed that these new models provide superior performance, making them the current state-of-the-art. Future developments could involve the usage of a fine-tuned biomedical LLM. Finally, future work in this study could focus on eCRF sections with an IAA below the 60% threshold to develop NLP tools and stop the regex approach, which has several limitations.

VI. Conclusions

In this work, we developed the entire clinical DB pipeline shared by fifteen IRCCS of the IVD. First, we defined a common CRF that contains clinical information that each Institute must investigate on its patients. Then we implemented NEMT, a software for semi-automatic extraction of items from eCRF. The core of NEMT is a BERT-based QABot, fine-tuned with the data collected by the IVD. While these data showed a moderate IAA, the results in the test set were relatively high (EM score = 78.1%, F1-score = 84.7%, LAcc = 0.834, MRR = 0.810) and were consistent with the literature, demonstrating the accuracy and correctness of the approach used. Moreover, NEMT outperformed LLMs such as ChatGPT and Vicuna in this specific topic. With this health informatics technology, we were able to populate a REDCap-based DB that in the future will contain data from thousands of patients across Italy, all evaluated with the same procedure. This effort paves the way for efficient extraction of clinical information and its adoption in new clinical research studies.

References

[1] H. Singh et al., "A qualitative study of hospital and community providers' experiences with digitalization to facilitate hospital-to-home transitions during the COVID-19 pandemic," PLOS ONE, vol. 17, no. 8, p. e0272224, Aug. 2022, doi: https://doi.org/10.1371/journal.pone.0272224.

[2] Y. Wang et al., "Clinical information extraction applications: A literature review," Journal of Biomedical Informatics, vol. 77, pp. 34–49, Jan. 2018, doi: https://doi.org/10.1016/j.jbi.2017.11.011.

[3] E. Joukes, A. Abu-Hanna, R. Cornet, and N. de Keizer, "Time Spent on Dedicated Patient Care and Documentation Tasks Before and After the Introduction of a Structured and Standardized Electronic Health Record," Applied Clinical Informatics, vol. 09, no. 01, pp. 046–053, Jan. 2018, doi: https://doi.org/10.1055/s-0037-1615747.

[4] A. Nigri et al., "Quantitative MRI Harmonization to Maximize Clinical Impact: The RIN–Neuroimaging Network," Frontiers in Neurology, vol. 13, Apr. 2022, doi: https://doi.org/10.3389/fneur.2022.855125.

[5] R. Grishman, "Information Extraction," IEEE Intelligent Systems, vol. 30, no. 5, pp. 8–15, Sep. 2015, doi: https://doi.org/10.1109/MIS.2015.68.

[6] K. Thompson. "Programming Techniques: Regular expression search algorithm," in Communications of the ACM, vol. 11, pp. 419 - 422, 1968.

[7] S. Wermter, E. Riloff, G. Scheler, "Connectionist, Statistical and Symbolic Approaches to Learning for Natural Language Processing," in Lecture Notes in Computer Science, 1996.

[8] E. Charniak. "Passing Markers: A Theory of Contextual Influence in Language Comprehension," in Cogn. Sci., vol. 7, pp. 171-190, 1983.

[9] A. Vaswani et al, "Attention Is All You Need," 2023, doi: https://doi.org/10.48550/arXiv.1706.03762.

[10] J. Devlin, M. Chang, K. Lee, K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," 2019, doi: https://doi.org/10.48550/arXiv.1810.04805.

[11] C. Raffel et al, "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," 2023, doi: https://doi.org/10.48550/arXiv.1910.10683.

[12] T. B. Brow. et al, "Language Models are Few-Shot Learners," 2020, doi: https://doi.org/10.48550/arXiv.2005.14165.

[13]P. Rajpurkar, J. Zhang, K. Lopyrev, P. Liang, "SQuAD: 100,000+ Questions for Machine Comprehension of Text," 2016, doi: https://doi.org/10.48550/arXiv.1606.05250.

[14]J. Lee et al., "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," Bioinformatics, vol. 36, no. 4, Sep. 2019, doi: https://doi.org/10.1093/bioinformatics/btz682.

[15]H. W. Chung et al, "Scaling Instruction-Finetuned Language Models," 2022, doi: https://doi.org/10.48550/arXiv.2210.11416.

[16]A. Conneau et al., "Unsupervised Cross-lingual Representation Learning at Scale," Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, doi: https://doi.org/10.18653/v1/2020.acl-main.747.

[17]S. Schweter, "Italian BERT and ELECTRA models". Zenodo, nov. 08, 2020. doi: https://doi.org/10.5281/zenodo.4263142.

[18]https://chat.openai.com/

[19]T. M. Buonocore, C. Crema, A. Redolfi, R. Bellazzi, E. Parimbelli, "Localizing in-domain adaptation of transformer-based biomedical language models," in Journal of Biomedical Informatics, vol. 144, pp. 104431, 2023, doi: https://doi.org/10.1016/j.jbi.2023.104431.

[20]C. Crema et al., "Advancing Italian biomedical information extraction with transformers-based models: Methodological insights and multicenter practical application," in Journal of Biomedical Informatics, vol. 148, pp. 104557, 2023, doi: https://doi.org/10.1016/j.jbi.2023.104557.

[21]G. Balikas, A. Krithara, I. Partalas, and G. Paliouras, "BioASQ: A challenge on large-scale biomedical semantic indexing and question answering," in Multimodal Retrieval in the Medical Domain. Cham, Switzerland: Springer, 2015, pp. 26–39, doi: https://doi.org/10.1007/978-3-319-24471-6_3

[22]P. Rajpurkar, R. Jia, P. Liang, "Know What You Don't Know: Unanswerable Questions for SQuAD," 2018, doi: https://doi.org/10.48550/arXiv.1806.03822.

[23]https://lmsys.org/blog/2023-03-30-vicuna/

[24]Hugo Touvron et al, "LLaMA: Open and Efficient Foundation Language Models," 2023, doi: https://doi.org/10.48550/arXiv.2302.13971.

[25]https://sharegpt.com/

[26]https://gdpr.eu/

[27]P. A. Harris, R. Taylor, R. Thielke, J. Payne, N. Gonzalez, and J. G. Conde, "Research electronic data capture (REDCap)—A metadata-driven methodology and workflow process for providing translational research informatics support," Journal of Biomedical Informatics, vol. 42, no. 2, pp. 377–381, Apr. 2019, doi: https://doi.org/10.1016/j.jbi.2008.08.010.

[28]P. A. Harris et al., "The REDCap consortium: Building an international community of software platform partners," Journal of Biomedical Informatics, vol. 95, no. 1, p. 103208, Jul. 2019, doi: https://doi.org/10.1016/j.jbi.2019.103208.

[29]Y. Wu et al, "Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation," 2016, doi: https://doi.org/10.48550/arXiv.1609.08144.

[30]P. Yadav et al., "Semantic Relations in Compound Nouns: Perspectives from Inter-Annotator Agreement". Studies in health technology and informatics, 245, 644–648, 2017.

[31]T. T. Dinh, T. P. Vo-Chanh, C. Nguyen, V. Q. Huynh, N. Vo, and H. D. Nguyen, "Extract antibody and antigen names from biomedical literature," BMC Bioinformatics, vol. 23, no. 1, Dec. 2022, doi: https://doi.org/10.1186/s12859-022-04993-4.

[32]K. Roberts, D. Demner-Fushman, "Annotating Logical Forms for EHR Questions," in Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), 2016, pp. 3772–3778.

[33]Peng, K., Yin, C., Rong, W., Lin, C., Zhou, D., & Xiong, Z. (2022). Named Entity Aware Transfer Learning for Biomedical Factoid Question Answering. IEEE/ACM transactions on computational biology and bioinformatics, 19(4), 2365–2376. https://doi.org/10.1109/TCBB.2021.3079339

[34]Xu, G., Rong, W., Wang, Y., Ouyang, Y., & Xiong, Z. (2021). External features enriched model for biomedical question answering. BMC bioinformatics, 22(1), 272. https://doi.org/10.1186/s12859-021-04176-7

[35]Tsatsaronis, G., Balikas, G., Malakasiotis, P. et al. An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. BMC Bioinformatics 16, 138 (2015). https://doi.org/10.1186/s12859-015-0564-6

[36]M. Boguslav, K. B. Cohen, "Inter-Annotator Agreement and the Upper Limit on Machine Performance: Evidence from Biomedical Natural Language Processing", Studies in health technology and informatics, 245, 298–302, 2017, doi: https://doi.org/10.3233/978-1-61499-830-3-298

[37]R. Azamfirei, S. R. Kudchadkar, and J. Fackler, "Large language models and the perils of their hallucinations," Critical Care, vol. 27, no. 1, Mar. 2023, doi: https://doi.org/10.1186/s13054-023-04393-x.

[38]T Kojima, S. S. Gu, M. Reid, Y. Matsuo, Y. Iwasawa, "Large Language Models are Zero-Shot Reasoners," 2023, doi: https://doi.org/10.48550/arXiv.2205.11916.

[39]C. Lamanna, L. Byrne, "Should Artificial Intelligence Augment Medical Decision Making? The Case for an Autonomy Algorithm," AMA Journal of Ethics, vol. 20, no. 9, pp. E902-910, Sep. 2018, doi: https://doi.org/10.1001/amajethics.2018.902.

[40]T. Lysaght, H. Y. Lim, V. Xafis, and K. Y. Ngiam, "AI-Assisted Decision-making in Healthcare," Asian Bioethics Review, vol. 11, no. 3, pp. 299–314, Sep. 2019, doi: https://doi.org/10.1007/s41649-019-00096-0.

[41]G. Spitale, N. Biller-Andorno, F. Germani, "AI model GPT-3 (dis)informs us better than humans," 2023, doi: https://doi.org/10.48550/arXiv.2301.11924.

[42]C. Crema, G. Attardi, D. Sartiano, and A. Redolfi, "Natural language processing in clinical neuroscience and psychiatry: A review," Frontiers in Psychiatry, vol. 13, Sep. 2022, doi: https://doi.org/10.3389/fpsyt.2022.946387.

[43]T.M., Buonocore et al., "A Rule-Free Approach for Cardiological Registry Filling from Italian Clinical Notes with Question Answering Transformers," in Artificial Intelligence in Medicine. AIME 2023. Lecture Notes in Computer Science, vol 13897. Springer, Cham. https://doi.org/10.1007/978-3-031-34344-5_19

[44]Islam, M.T., Zhou, Z., Ren, H. et al. Revealing hidden patterns in deep neural network feature space continuum via manifold learning. Nat Commun 14, 8506 (2023). https://doi.org/10.1038/s41467-023-43958-w