Claudio Crema\textsuperscript{1}, Federico Verde\textsuperscript{1}, Pietro Tiraboschi\textsuperscript{1}, Camillo Marra\textsuperscript{1}, Andrea Arighi\textsuperscript{1}, Silvia Fostinelli\textsuperscript{1}, Guido Maria Giuffrè\textsuperscript{1}, Vera Pacoova Dal Maschio\textsuperscript{1}, Federica L’abbate\textsuperscript{1}, Federica Solca\textsuperscript{1}, Barbara Poletti\textsuperscript{1}, Vincenzo Silani\textsuperscript{1}, Emanuela Rotondo\textsuperscript{1}, Vittoria Borrazzi\textsuperscript{1}, Roberto Vimercati\textsuperscript{1}, Valeria Crespaldi\textsuperscript{1}, Emanuela Ingusco\textsuperscript{1}, Massimo Filippi\textsuperscript{1}, Francesca Caso\textsuperscript{1}, Alessandra Maria Rosati\textsuperscript{1}, Davide Quaranta\textsuperscript{1}, Giuliano Binetti\textsuperscript{1}, Ilaria Pagnoni\textsuperscript{1}, Manuela Morreale\textsuperscript{1}, Francesca Burgio\textsuperscript{1}, Michelangelo Stanzani Maserati\textsuperscript{1}, Sabina Capellari\textsuperscript{1}, Matteo Pardini\textsuperscript{1}, Nicola Girtler\textsuperscript{1}, Federica Piras\textsuperscript{1}, Fabrizio Piras\textsuperscript{1}, Stefania Lalli\textsuperscript{1}, Elena Perdixi\textsuperscript{1}, Gemma Lombardi\textsuperscript{1}, Sonia Di Tella\textsuperscript{1}, Alfredo Costa\textsuperscript{1}, Marco Capelli\textsuperscript{1}, Cira Fundarò\textsuperscript{1}, Marina Manera\textsuperscript{1}, Cristina Muscio\textsuperscript{1}, Elisa Pellencin\textsuperscript{1}, Raffaele Lodi\textsuperscript{1}, Fabrizio Tagliavini\textsuperscript{1}, and Alberto Redolfi\textsuperscript{1}

\textsuperscript{1}Affiliation not available

December 22, 2023

Abstract

The advent of computerized medical recording systems in healthcare facilities has made data retrieval tasks easier, compared with 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 database. NEMT core is a Question Answering Bot (QABot) based on a modern NLP model, fine-tuned using thousands of examples produced from IVD centers. Detailed descriptions of the process for defining a common minimum dataset, the Inter-Annotator Agreement calculated on clinical documents, and NEMT results, are provided. The best QABot performance in terms of Exact Match (EM) and F1-score (78.1\% and 84.7\%) outperforms ChatGPTv3.5 (68.9\% and 52.5\%). NEMT represents an efficient tool that paves the way for medical information extraction and exploitation for new research studies.
Abstract— The advent of computerized medical recording systems in healthcare facilities has made data retrieval tasks easier, compared with 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 (VDI), 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 database. NEMT core is a Question Answering Bot (QABot) based on a modern NLP model, fine-tuned using thousands of examples produced from IVD centers. Detailed descriptions of the process for defining a common minimum dataset, the Inter-Annotator Agreement calculated on clinical documents, and NEMT results, are provided. The best QABot performance in terms of Exact Match (EM) and F1-score (78.1% and 84.7%) outperforms ChatGPTv3.5 (68.9% and 52.5%). 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.

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


This work was partially funded by the National funding of Italian Ministry of Economy and Finance (CCR-2017-23696978) and by the National funding of the Italian Ministry of Health under the framework of the grant ISTITUTI NAZIONALI VIRTUALI (RCR 2020–23670067 and RCR 2021–23671214) and in the framework of the grant PROGETTO RETE NIN 2022 (RCR-2022-23683294).

(Corresponding author: C. Crema)

Claudio Crema and Alberto Redolfi are with the Department of Neuroinformatics (e-mail: ccrema@fatebenefratelli.eu, arendolfi@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 Illaria Pagnoni is with the Neuropsyscology Unit (e-mail: pagonnai@fatebenefratelli.eu), all adherent 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: l.verde@auxologico.it, f.solca@auxologico.it, b.poletti@auxologico.it, v.silani@iuliana.com).

Pietro Tiraboschi, Valeria Crepaldi, Emanuela Inguscio, and Elsa Pelliccioni are with the Division of Neurology (e-mail: pietro.tiraboschi@istituto-besta.it, valeria.crepaldi@istituto-besta.it, elmas.pelliccioni@istituto-besta.it), and Fabrizio Tagliavini is with the Scientific Directorate, all adherent to the Fondazione IRCCS Istituto Neurologico Carlo Besta, Milan, Italy (e-mail: f.tagliavini@istituto-besta.it, cirmia@fondazioneirccss.org).

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).

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

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

Francesca Burgio is with the Neuropsychology Department of the IRCCS San Camillo Hospital, Venice, Italy (e-mail: francesca.burgio@hsanmario.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.stanzanimasera@unibo.it, sabina.capellari@unibo.it, raffaele.lodi@unibo.it).

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

Cristina Muscio was with the Division of Neurology of the Fondazione IRCCS Istituto Neurologico Carlo Besta, Milan, Italy (e-mail: c.r.muscio@asst-bogvet.it).

Camillo Marra, Guido Maria Giuffre, Federica L’Abbate, and Alessandra Maria Rosati are with the UO Menta Clinica 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, federica.labbate@gmail.com, alessandrarma.rosati@policlinicogemelli.it). Andrea Arighi, Emanuela Rotondo, Vittoria Borracci, and Roberto Vimercati are with the Neuroneurodegenerative Diseases Unit of the Fondazione IRCCS Ca’ Granda Ospedale Maggiore Policlinico, Milan, Italy (e-mail: andre.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).
I. INTRODUCTION

Digital technologies are becoming more and more pervasive in healthcare facilities [1], particularly in those like Scientific Institutes for Research, Hospitalization and Healthcare (named IRCCS, from the Italian acronym of these organizations) 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 practice of clinicians (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. 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 was founded in 2017 by the Ministry of Health to stimulate collaboration among IRCCSs, to foster the dissemination of information on clinical-scientific activity, facilitating scientific and technological research and training, to coordinate internationally relevant actions aimed at increasing the prominence and competitiveness of the field, and to transfer research results into the industrial supply chain (technology transfer). RIN is composed by the so-called 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; the complete list can be found in Supplementary materials. Out of the sixteen institutes comprised in the IVD, one 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 show a lack of adaptability when it comes to expounding upon documents that are written in free text format. Complex documents require increasingly complicated algorithms (e.g., getting information from a free text can be extremely challenging) [39]: 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 step 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. However, 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 biomedical. 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. The most notorious is ChatGPT [18], released at the end of 2022; after that, other open-source LLMs have been released (e.g.: Vicuna, available at https://huggingface.co/lmsys/vicuna-1.3b-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 text (called “prompt”) or in the
original training dataset and generate new information. However, LLMs typically present some critical issues:

- They can only be accessed online, 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 General Data Protection Regulation (GDPR) [26];
- They are trained on broad topics corpora; thus, they could struggle when it comes to very specific topics.

This work focuses on the clinical and technological decisions that led from the conceptualization of a harmonized CRF within the IVD, to the development of a pipeline for the extraction of relevant information from eCRFs. While the definition of a common eCRF structure among fifteen independent hospitals all over Italy is a notable result, and the implementation of a CRF-to-DB automated pipeline is also a remarkable achievement, the true technological improvement of this work is the creation of an NLP model able to extract data from unstructured clinical documents. In the following paragraphs 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:

1. IVD eCRF consensus: harmonization and crafting of the eCRF;
2. Information extraction: semi-automated data extraction from the clinical document;
3. 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 2 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). Whenever possible, the CRF was directly incorporated in the Institutes’ electronic medical charts. It is important to note that the IRCCSs 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 records. In some Institute, 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 in the concerned Institute so that all information was collected and recorded 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, who was the leader of the neuropsychology working group. The complete list of items of the two eCRF consensuses can be found in the Supplementary materials.

Information Extraction

**NEMT**

The goal of this task was to develop a tool able to automatically extract CRF Consensus items from the IVD centers’ eCRFs. The main difficulty was the lack of a common structured template for the IVD. After extracting the items from the eCRF, 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”. We have set up a local instance of NEMT for each IVD center:

1. The patient is visited by the clinician, who creates an eCRF in PDF format using the hospital’s electronic medical record. The clinical document is then converted to text by NEMT, with headers, footers, and unnecessary information removed;
2. The items of the consensus are then extracted using the text-mining algorithm, which combines regexes for well-defined data (e.g., Italian fiscal-code) and NLP for open questions (e.g., diagnostic hypotheses). 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;
3. Finally, the data are converted in a well-defined structure and sent to the remote REDCap. 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.
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 IE task is semi-automated: items extracted 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 [38], thus a human quality check is always advised. Indeed, some human intervention 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, the AI tool developed in NEMT 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 becoming part of the clinical decision-making process, speeding up the whole tasks and contributing to lower the cost of medicine [36] [37] through significant optimizations.

**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 developed in [19] and [20]:

- BioBIT (available at https://huggingface.co/IVN-RIN/bioBIT), a BERT-based checkpoint pre-trained on the Italian translated version of the original BioBERT corpus;
- MedBIT (available at https://huggingface.co/IVN-RIN/medBIT) and MedBIT-r3-plus (available at 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 by three parts: a context, i.e., human-written text containing the information that the model will elaborate to retrieve the required data; a human-written 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 by 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) is, by definition, a measure that expresses the extent to which two or more annotators agree on a decision, in this particular case when answering questions. Evaluating IAA is a crucial step in ensuring the quality of labeled datasets. Determining which IAA values are considered high enough is complicated and depends on many factors. Nevertheless, a value below 20% is considered poor, between 20 and 40% fair, between 40 and 60%
moderate, while values above 60% are considered good, and above 80% excellent. IAA can be calculated in several ways; we decided to use Exact Match (EM) and F1-scores, in order to have values comparable to the algorithm performance. EM-score, a number ranging from 0 to 1, is the ratio of answers that perfectly match between gold standard and annotator. F1-score is calculated by comparing every answer with the respective gold standard one, counting the shared tokens (see Supplementary materials for an accurate description of what tokens are), and thus calculating True Positives (TPs), False Positives (FPs), and False Negatives (FNs). With these values it is possible to calculate Precision and Recall, while F1-score is the harmonic mean of Precision and Recall; the complete formulas are presented in the Supplementary materials. To calculate the F1-score on several questions, we used Macro-F1, obtained by averaging the single F1-scores. This option gives the same weight to every answer, regardless of the original number of tokens.

IAA for real-world scenarios presents a wide range of values. Previous works on different corpora report variable data, as 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% (e.g.: concept normalization) to 88% (e.g.: temporal expressions). It appears that the IAA score depends on the complexity of the problem, and it is not uncommon to have relatively low values even when annotations are performed by highly qualified clinical staff.

To calculate IAA, we randomly selected five of the fifteen centers of the IVD. Every center 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 reference answers (gold standard) are the annotations of the clinician coming from 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 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 by about 60k QA examples in total;
- Neuropsychological IVD dataset, composed by about 36k QA examples in total.

We exploited these datasets to train the BERT-based models mentioned above. In order to increase the size of the datasets, and thus to possibly enhance the model performance, we combined IVD datasets with the following ones:

- BioASQ SQuAD [21]: Biological Question Answering (BioASQ) is a research initiative, organized as a series of challenges in the field of biomedical semantic indexing and QA. This dataset is composed by datasets from versions 4, 5, and 6 of the challenge, for a total of about 8 thousand examples. The original datasets are in English, so we translated them using the Google’s neural machine translation system [29]. These datasets are related to biomedical concepts, so using them for training could increase the capability of the models to elaborate information related to a clinical context;
- SQuAD: not all questions of the Consensus CRF require a deep medical experience to be answered; some of them can be handled by understanding the grammatical structure of the context. For these reasons, we decided to include the SQuAD dataset; although it is not biomedical, it contains examples of generic questions, and thus it can help to improve the efficacy of the models.

We ran 3 fine-tuning experiments:

1. In the first, we fine-tuned Bio/MedBIT with the IVD dataset;
2. In the second, we fine-tuned Bio/MedBIT with a dataset created by merging the IVD dataset with BioASQ SQuAD;
3. In the third, we fine-tuned Bio/MedBIT with a dataset created by merging the IVD dataset plus BioASQ SQuAD and generic SQuAD 2.0.

Large Language Models tests

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-3.5 architecture (depending on the version). 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 exploited to perform test on manually 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. Details related to this task are discussed in the Database Section. 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 supports online and offline data capture for research studies and operations. It also includes robust data validation rules that enhance data quality and provide researchers with reliable and error-free data for analysis. REDCap provides advanced security measures that ensure data confidentiality in compliance with regulatory standards, e.g., GDPR. REDCap also provides APIs that allow interaction with external software; by using this mechanism, it is 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 clinical practitioners.

III. RESULTS

eCRF Consensus

The result of this work was the definition of two lists of items, one clinical and one neuropsychological. This list represents the minimum dataset that every eCRF had to provide, regardless of the institute in which the patient was assessed. The complete lists can be found in the Supplementary materials.

Information Extraction

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, to be consistent with performance of fine-tuning. 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 rationale is that, in case of very different results from different sections, we could apply the QA Bot 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 QA model would only work on items of the Objective exam section. 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 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 test, 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, 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

Large Language Models

For research purposes, two LLMs, namely ChatGPT (3 August 2023 version) and Vicuna, were tested on the same task of the fine-tuned QA Bot. The tests were run on 10% of the three test sets, and then the EM and F1-scores were calculated and compared (results are shown in TABLE IV), although 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, would 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

<table>
<thead>
<tr>
<th>Center</th>
<th>EM score [%]</th>
<th>F1-score [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No answer</td>
<td>Text answer</td>
</tr>
<tr>
<td>Center 1</td>
<td>78.1 ± 9.4</td>
<td>29.2 ± 14.8</td>
</tr>
<tr>
<td>Center 2</td>
<td>68.2 ± 15.5</td>
<td>40.9 ± 13.0</td>
</tr>
<tr>
<td>Center 3</td>
<td>87.3 ± 8.9</td>
<td>33.8 ± 14.4</td>
</tr>
<tr>
<td>Center 4</td>
<td>78.1 ± 9.7</td>
<td>39.9 ± 8.9</td>
</tr>
<tr>
<td>Center 5</td>
<td>87.4 ± 7.2</td>
<td>34.9 ± 11.1</td>
</tr>
<tr>
<td>Overall</td>
<td>79.9 ± 11.8</td>
<td>35.7 ± 12.0</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Section</th>
<th>EM score [%]</th>
<th>F1-score [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Section</td>
<td>No answer</td>
<td>Text answer</td>
</tr>
<tr>
<td>Clinical history</td>
<td>38.8 ± 24.1</td>
<td>12.7 ± 13.5</td>
</tr>
<tr>
<td>Risk factors</td>
<td>74.6 ± 20.4</td>
<td>0.8 ± 1.3</td>
</tr>
<tr>
<td>Comorbidity</td>
<td>23.3 ± 38.3</td>
<td>10.0 ± 5.0</td>
</tr>
<tr>
<td>Neurological examination</td>
<td>22.6 ± 20.2</td>
<td>28.9 ± 15.5</td>
</tr>
<tr>
<td>Drugs</td>
<td>28.3 ± 24.7</td>
<td>35.8 ± 21.8</td>
</tr>
</tbody>
</table>

### TABLE III
TRAINING SCORES FOR THE THREE STARTING CHECKPOINTS (BioBIT, MedBIT, MedBIT-r3) ON THREE DIFFERENT DATASETS (IVD, IVD + BioASQ, IVD + BioASQ + SQuAD)

<table>
<thead>
<tr>
<th>Starting model</th>
<th>Fine-tuning dataset</th>
<th>EM score [%]</th>
<th>F1-score [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No answer</td>
<td>Text answer</td>
<td>Overall</td>
</tr>
<tr>
<td>BioBIT</td>
<td>IVD</td>
<td>97.5</td>
<td>52.2</td>
</tr>
<tr>
<td></td>
<td>IVD + BioASQ</td>
<td>97.7</td>
<td>66.5</td>
</tr>
<tr>
<td></td>
<td>IVD + BioASQ + SQuAD</td>
<td>97.2</td>
<td>63.3</td>
</tr>
<tr>
<td>MedBIT</td>
<td>IVD</td>
<td>96.9</td>
<td>50.8</td>
</tr>
<tr>
<td></td>
<td>IVD + BioASQ</td>
<td>97.8</td>
<td>64.4</td>
</tr>
<tr>
<td></td>
<td>IVD + BioASQ + SQuAD</td>
<td>97.9</td>
<td>63.9</td>
</tr>
<tr>
<td>MedBIT-r3</td>
<td>IVD</td>
<td>98.6</td>
<td>82.4</td>
</tr>
<tr>
<td></td>
<td>IVD + BioASQ</td>
<td>96.9</td>
<td>65.8</td>
</tr>
<tr>
<td></td>
<td>IVD + BioASQ + SQuAD</td>
<td>98.2</td>
<td>63.5</td>
</tr>
</tbody>
</table>

### TABLE IV
RESULTS OF IE EXECUTED BY MEANS OF LLMs ChatGPT and Vicuna

<table>
<thead>
<tr>
<th>Starting model</th>
<th>Test set</th>
<th>EM score [%]</th>
<th>F1-score [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No answer</td>
<td>Text answer</td>
<td>Overall</td>
</tr>
<tr>
<td>ChatGPT (v3.5)</td>
<td>IVD</td>
<td>76.6</td>
<td>29.4</td>
</tr>
<tr>
<td></td>
<td>IVD + BioASQ</td>
<td>84.3</td>
<td>58.9</td>
</tr>
<tr>
<td></td>
<td>IVD + BioASQ + SQuAD</td>
<td>87.2</td>
<td>48.9</td>
</tr>
<tr>
<td>Vicuna (v13b)</td>
<td>IVD</td>
<td>74.6</td>
<td>23.76</td>
</tr>
<tr>
<td></td>
<td>IVD + BioASQ</td>
<td>88.9</td>
<td>48.9</td>
</tr>
<tr>
<td></td>
<td>IVD + BioASQ + SQuAD</td>
<td>86.2</td>
<td>38.9</td>
</tr>
</tbody>
</table>
**Database**

At the time of writing, NEMT was being used in the IVD to process and store items from two hundred eCRFs.

**IV. Discussion**

**eCRF Consensus**

From a clinical point of view, the most important achievement of the clinical and neuropsychological eCRF is the standardization 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 homogenous 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.

Innovative and relevant aspects of the clinical eCRF of IVD include the systematic collection of data, the description of the patient’s and care-giver/informant perspectives in the history of neurological symptoms, the inclusion of a detailed family history and medical comorbidities, and the completeness of the neurological examination.

**Information Extraction**

**Inter-Annotator Agreement**

In terms of IAA performance, the EM scores on empty answers are higher than for non-empty answers, as it is usually easier for a rater to recognize a missing answer to a question than a correct answer. An overall EM-score of 56% means that, on average, the answers of all annotators match perfectly in 1 out of 2 cases. In addition, the overall EM-score for empty answers is close to 80%, which means that 4 out of 5 annotators agree that a question cannot be answered. A macro F1-score of about 79% is a good result. This 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, these results illustrate 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, the first three (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 the interpretability of the single rater, results of objective exams, and drugs even more, are objectively defined and thus easier to annotate in a correct way.

**Dataset creation, training, and fine-tuning of NEMT**

As expected, the EM scores are higher for empty answers than for text answers. The best results are provided by the BioBIT model fine-tuned with the IVD and BioASQ datasets merged together, and this pooled dataset provides the best scores for the other two starting checkpoints as well. This is likely because all the pooled questions were 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 (EM difference ranging from -0.4% to -2.5%, F1-score difference from 0% to -2.6%). 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. It is interesting 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 [33]. 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 choice of model architecture, and other factors.

By incorporating the developed QA models, NEMT was able to extract information from eCRFs from fifteen different institutes with high performance. These data were processed and efficiently stored safely 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. 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, namely 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 [34]), even if this was explicitly forbidden in the prompt. This was seen during the computation of the EM score (29.4% for ChatGPT and 23.8% for Vicuna), where a single different character resulted in a score of 0. 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 [35]. 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, models pre-trained on corpora of relevant domains can perform better with appropriate fine-tuning, regardless of having a much smaller number of parameters.

**Database**

As a result of our work, we have created an efficient IVD pipeline to store harmonized data from fifteen IRCCSs. These data could be used to test new scientific hypotheses and understand the onset mechanisms of AD and other forms of dementia. The data could also be used in future clinical research studies.

**V. LIMITATION AND FUTURE WORK**

The present work had a few limitations. First, the created QABot models were fine-tuned to a narrow topic specific for dementia. Thus, applying them to other medical topics would require new training from scratch. Then, even though the fine-tuning results were relatively high, the IAA values showed that the agreement between the annotators 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 Alzheimer’s disease. 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 of this work 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. 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 IRCCSs 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 of 78.1% and F1-score of 84.7%) 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**


