ARGO, Automatic Record Generator for Oncology: a natural language process-based tool to capture pathology features from onco-hematological reports

Gian Maria Zaccaria, Vito Colella, Simona Colucci, Felice Clemente, Fabio Pavone, Maria Carmela Vegliante, Flavia Esposito, Giuseppina Opinto, Antonio Negri, Giacomo Loseto, Carla Minoia, Bernardo Rossini, Vito Angiulli, Angela Maria Quinto, Laura Schirosi, Anna Scattone, Alfredo Zito, Luigi Alfredo Grieco, Attilio Guarini, Sabino Ciavarella

Submitted to: Journal of Medical Internet Research
on: January 22, 2021

Table of Contents

ARGO, Automatic Record Generator for Oncology: a natural language process-based tool to capture pathology features from onco-hematological reports

Gian Maria Zaccaria[1] ME, PhD; Vito Colella[2] MSc; Simona Colucci[2] PhD; Felice Clemente[1] MSc; Fabio Pavone[1] PhD; Maria Carmela Vegliante[1] PhD; Flavia Esposito[1, 3] PhD; Giuseppina Opinto[1] PhD; Antonio Negri[1] MSc; Giacomo Loseto[1] MD; Carla Minoia[1] MD; Bernardo Rossini[1] MD; Vito Angiulli[4] PhD; Angela Maria Quinto[1] MD; Laura Schirosi[5] MD; Anna Scattone[5] MD; Alfredo Zito[5] MD; Luigi Alfredo Grieco[2] Prof Dr; Attilio Guarini[1] MD; Sabino Ciavarella[1] MD, PhD

[1]Hematology and Cell Therapy Unit, IRCCS Istituto Tumori 'Giovanni Paolo II' Bari IT
[2]Department of Electrical and Information Engineering, Politecnico of Bari Bari IT
[3]Department of Mathematics, University of Bari Aldo Moro Bari IT
[4]Clinical Engineering Unit, IRCCS Istituto Tumori 'Giovanni Paolo II' Bari IT
[5]Pathology Department, IRCCS Istituto Tumori 'Giovanni Paolo II' Bari IT

Corresponding Author:
Gian Maria Zaccaria ME, PhD
Hematology and Cell Therapy Unit, IRCCS Istituto Tumori 'Giovanni Paolo II'
Viale Orazio Flacco, 65
Bari
IT

Abstract

Background: The unstructured nature of medical data from Real-World (RW) patients and the scarce accessibility for researchers to integrated systems restrain the use of RW information for clinical and translational research purposes. Natural Language Processing (NLP) might help in transposing unstructured reports in electronic health records (EHR), thus prompting their standardization and sharing.

Objective: We aimed at designing a tool to capture pathological features directly from hemo-lymphopathology reports and automatically record them into electronic case report forms (eCRFs).

Methods: We exploited Optical Character Recognition and NLP techniques to develop a web application, named ARGO (Automatic Record Generator for Oncology), that recognizes unstructured information from diagnostic paper-based reports of diffuse large B-cell lymphomas (DLBCL), follicular lymphomas (FL), and mantle cell lymphomas (MCL). ARGO was programmed to match data with standard diagnostic criteria of the National Institute of Health, automatically assign diagnosis and, via Application Programming Interface, populate specific eCRFs on the REDCap platform, according to the College of American Pathologists templates. A selection of 239 reports (n. 106 DLBCL, n.79 FL, and n. 54 MCL) from the Pathology Unit at the IRCCS - Istituto Tumori “Giovanni Paolo II” of Bari (Italy) was used to assess ARGO performance in terms of accuracy, precision, recall and F1-score.

Results: By applying our workflow, we successfully converted 233 paper-based reports into corresponding eCRFs incorporating structured information about diagnosis, tissue of origin and anatomical site of the sample, major molecular markers and cell-of-origin subtype. Overall, ARGO showed high performance (nearly 90% of accuracy, precision, recall and F1-score) in capturing identification report number, biopsy date, specimen type, diagnosis, and additional molecular features.

Conclusions: We developed and validated an easy-to-use tool that converts RW paper-based diagnostic reports of major lymphoma subtypes into structured eCRFs. ARGO is cheap, feasible, and easily transferable into the daily practice to generate REDCap-based EHR for clinical and translational research purposes.

(JMIR Preprints 22/01/2021:27295)
DOI: https://doi.org/10.2196/preprints.27295

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Original Paper

ARGO, Automatic Record Generator for Oncology: a natural language process-based tool to capture pathology features from onco-hematological reports.

Authors

Gian Maria Zaccaria1, Vito Colella2, Simona Colucci2, Felice Clemente1, Fabio Pavone1, Maria Carmela Vegliante1, Flavia Esposito1,3, Giuseppina Opinto1, Antonio Negri1, Giacomo Loseto1, Carla Minoia1, Bernardo Rossini1, Vito Angiulli4, Angela Maria Quinto1, Laura Schirosi5, Anna Scattone5, Alfredo Zito5, Luigi Alfredo Grieco2, Attilio Guarini1, and Sabino Ciavarella1

1 Hematology and Cell Therapy Unit, IRCCS Istituto Tumori 'Giovanni Paolo II', Bari, Italy
2 Department of Electrical and Information Engineering, Politecnico of Bari, Bari, Italy
3 Department of Mathematics, University of Bari Aldo Moro, Bari, Italy
4 Clinical Engineering Unit, IRCCS Istituto Tumori 'Giovanni Paolo II', Bari, Italy
5 Pathology Department, IRCCS Istituto Tumori 'Giovanni Paolo II', Bari, Italy

Abstract

Background. The unstructured nature of medical data from Real-World (RW) patients and the scarce accessibility for researchers to integrated systems restrain the use of RW information for clinical and translational research purposes. Natural Language Processing (NLP) might help in transposing unstructured reports in electronic health records (EHR), thus prompting their standardization and sharing. **Objective**. We aimed at designing a tool to capture pathological features directly from hemo-lymphopathology reports and automatically record them into electronic case report forms (eCRFs). **Method**. We exploited Optical Character Recognition and NLP techniques to develop a web application, named ARGO (Automatic Record Generator for Oncology), that recognizes unstructured information from diagnostic paper-based reports of diffuse large B-cell lymphomas (DLBCL), follicular lymphomas (FL), and mantle cell lymphomas (MCL). ARGO was programmed to match data with standard diagnostic criteria of the National Institute of Health, automatically assign diagnosis and, via Application Programming Interface, populate specific eCRFs on the REDCap platform, according to the College of American Pathologists templates. A selection of 239 reports (n. 106 DLBCL, n.79 FL, and n. 54 MCL) from the Pathology Unit at the IRCCS - Istituto Tumori “Giovanni Paolo II” of Bari (Italy) was used to assess ARGO performance in terms of accuracy, precision, recall and F1-score. **Results**. By applying our workflow, we successfully converted 233 paper-based reports into corresponding eCRFs incorporating structured information about diagnosis, tissue of origin and anatomical site of the sample, major molecular markers and cell-of-origin subtype. Overall, ARGO showed high performance (nearly 90% of accuracy, precision, recall and F1-score) in capturing identification report number, biopsy date, specimen type, diagnosis, and additional molecular features. **Conclusions**. We developed and validated an easy-to-use tool that converts RW paper-based diagnostic reports of major lymphoma subtypes into structured eCRFs. ARGO is cheap, feasible, and easily transferable into the daily practice to generate REDCap-based EHR for clinical and translational research purposes.

Keywords: real world data collection, natural language processing, optical character recognition, electronic case report forms, redcap, pathology, clinical data.

Introduction

Over the last few years, the complexity of clinical and biological data for a proper diagnosis and prognostication of onco-hematological diseases remarkably increased, especially in the field of lymphomas (1,2). In parallel, novel therapeutics found continue approvals from large, controlled trials, but missed parallel validation in the Real-World (RW) setting (3). This major controversy claims for an urgent improvement of the capability to collect and share RW data with the final goal to support clinical and translational research (4). Frequently, RW data are derived from fragmented sources as medical registries, electronic records, computerized patient order entries, individual databases, paper notes, as well as monocentric bio-banking-related annotations. Moreover, the common dearth of specialized data-entry professionals and the uneasy accessibility to data-extraction systems for most physicians accentuate the need of tools that facilitate the process of health data recording (3).

Natural language processing (NLP) is an useful technique to extract essential unstructured data from diagnostic and prognostic notes (6–10) in onco-hematology (11–16), as well as REDCap is recognized as a platform of electronic case report forms (eCRFs) enabling rapid, high-quality and standardized annotation of data (17,18). On the other hand, Optical Character Recognition (OCR) is a technology able to convert paper-based reports into digital forms to be further structured – possibly thorough NLP techniques – in electronic health records (EHR), thus overcoming the need of integration between systems (19,20).

Here, we described the design and functions of ARGO (Automatic Record Generator for Oncology), a web application leveraging the combination of an “ad-hoc” NLP-based tool with REDCap to convert RW pathological reports in standardized eCRFs for automatic data collection.

We applied the framework onto a monocentric set of Non-Hodgkin Lymphomas (NHL) and validated its functionality, performance, and feasibility in the daily practice.

Methods

Data collection

We selected 239 histopathology paper-based reports collected between 2014 and 2020 at the Pathology Unit of the IRCCS - Istituto Tumori 'Giovanni Paolo II' in Bari (Italy). The reports were conclusive for diagnoses of n. 106 Diffuse Large B-Cell Lymphoma (DLBCL), n. 79 Follicular Lymphoma (FL), and n. 54 Mantle Cell Lymphoma (MCL), while for n. 19 reports the diagnosis was field cases. A unique identification code (ID) was assigned to each report. According to the diagnostic criteria for each lymphoma subtype, reports included immunohistochemical (IHC) results obtained from lymph-node (LN), extra-nodal (EN), bone marrow (BM) or peripheral blood (PB) specimens. Qualitative and quantitative information for IHC markers including MYC, BCL2, BCL6, Cluster of Differentiation (CD)10, CD20, Cyclin-D1 were reported. Some reports also included molecular data from fluorescent in situ hybridization (FISH) analysis, while some reports included either FISH results or the level tumor cell infiltration as addendum. For DLBCL, molecular classification according to the cell-of-origin (COO) estimated by the Hans algorithm was also included (21). Ki-67 proliferation index was also reported as quantitative value ranging from 5 to 100%.

Automated detection of relevant terms in paper-based reports

We aimed this phase of the workflow at developing a web application to automate the detection of relevant terms to be extracted from the text fields of paper-based pathology reports. ARGO exploits OCR (22) and NLP (23) techniques to i) convert images of reports into text and detect relevant words in the text based on an “ad-hoc” thesaurus.

The conversion from image to text has been implemented in Tesseract OCR© version 4.1.1-rc2-20-

g01fb. To improve conversion performance, each pathology report was firstly converted from pdf to image through Poppler library, version 0.26.5 called inside Python app thank pdf2image version 1.13.1. Then, the image is translated in a grey scale of 8 bits (from 0 to 255 levels of grey), according to the formula (24):

$$dst(x,y) = \begin{cases} maxval, if src(x,y) > thresh \\ 0, otherwise \end{cases}$$
$$where\ thresh = 120$$

Image transformation was developed in Python by OpenCV© software, v. 4.2.0.
In ARGO, NLP techniques were adopted to automatically extract terms relevant for the disease diagnosis to be transferred into the digitalized eCRFs. In particular, a set of NLP regular expressions was matched to the text in order to find information including diagnosis, date of the report, report ID, type of the specimen, execution of BM biopsy, IHC, and FISH analyses, as well as quantitative and qualitative data of selected IHC markers (MYC, BCL2, BCL6, CD10, CD20, Cyclin-D1), COO subtypes and Ki-67 proliferation index. The disease nomenclature was assigned based on the highest match between the pattern of detected biomarkers in each report and a reference pattern, as reported in the “Hematopoietic and Lymphoid Neoplasm Coding Manual guidelines from the Surveillance, Epidemiology, and End Result (SEER) program” of the National Institute of Health (25). The final diagnosis nomenclature was referred to the “International Statistical Classification of disease and related health problems 10th revision” (ICD10, version 2019, World Health Organization) (26). Communication between ARGO and SEER official servers was dealt via Application Programming Interface (API).
ARGO was developed in Flask©, version 1.1.2, the webserver was an Oracle© Linux Server 7.8 with kernel 4.14.35-1902.303.5.3.el7uek.x86_64. We used MariaDB© 5.5.68 as database. NLP algorithms were developed in Python 3.6.8 with various regular expressions. Translation from English to Italian language was dealt via API tool MyMemory©, version 3.5.0. In order to increase the detectability of biomarkers in the reports we used Python to build three thesauri (Multimedia Appendix 1 Source code S1, Multimedia Appendix 1 Table S1).

Data-mapping and automatic population of eCRFs

For a systematic collection of the diagnostic variables in this study, we designed dedicated eCRFs on REDcap (17,18). eCRFs were suited to the synoptic templates provided and approved by the College of American Pathologists. We referred to DLBCL, FL, and MCL templates (27,28). The data-mapping between ARGO and the eCRFs was performed by providing the relevant data fields from the REDCap dictionary to the NLP code (Multimedia appendix 1 Table S2). Finally, we used API technology for the automatic data entry and final upload of the information of interest into the eCRFs.

Validation metrics

ARGO performance, regarded as the level of consistency between data included in the original pathology reports and those automatically transferred into eCRFs, was assessed in terms of accuracy, precision, recall and F1 score, as reported (29). To calculate each measure, we defined as i) true-positive those cases in which ARGO detected correctly the expected variables; ii) false-positive those cases in which ARGO detected variables even if not available; iii) true-negative those cases in which ARGO did not detect an unavailable variable; and iv) false-negative those cases in which ARGO failed in detecting an available variable.

Results

Electronic data collection workflow

Based on the template currently adopted at the Pathology Unit of the IRCCS-Istituto Tumori "Giovanni Paolo II", each histopathology report included several data organized in four main sections: i) biopsy date and ID report number; ii) demographical patient information; iii) specimen characteristics; and iv) biomarker and diagnosis description (Figure 1A).
The first step of our workflow consisted in the advantageous transformation of each paper-based report into an image file (.jpg extension) by using a common digital scanner. Thus, each report was uploaded on the ARGO application, which saved structured text into a support database, retrieved all the relevant data from the text, and transferred them directly into dedicated eCRFs. 233 out of 239 reports were successfully converted in eCRF records. Figure 1B depicts the main sections of each eCRF, which included both "demographic" and "disease" modules, in a way consistent with the sections in the corresponding original paper report. Six failures were recorded for those reports either built on alternative templates or longer more than 1 paper page or with low optical quality. A video demonstrates the ARGO functionality in Multimedia Appendix 2.

Characteristics of data retrieved from diagnostic reports

Among the 239 paper-based reports collected for the study, n. 106, n. 79, and n. 54 were respectively conclusive for a diagnosis of DLBCL, FL, and MCL (Figure 2A). Overall, n. 117 diagnostic specimens were obtained from LN, n. 73 were EN, and n. 34 from BM, n. 2 from PB, and for n. 12 cases this information was not available (Figure 2B). In 85% of cases, a matched bone marrow biopsy was available (Figure 2C). Results from IHC staining for MYC, BCL2, BCL6, CD10, CD20, and Cyclin-D1 were available in 229 out of 239 cases and included a qualitative (positive/negative) assessment for the most relevant biomarkers (Figure 2D-E). A FISH analysis (for MYC, BCL2, BCL6 or Cyclin-D1) appeared in the 30% of reports (Figure 2F), whereas COO categorization was reported in nearly 20% of cases (Figure 2G). Of note, 187/239 reports included the quantitative value of the Ki-67. Among these, 54 reported a value lower than 30% (Figure 2H).

Validation

Overall, ARGO detected 92,121 terms of interest and successfully generated EHR for the 97% of the processed histopathology reports. Figure 3(A-H) shows the post-hoc validation of ARGO performance for all the studied data fields. Concerning the "diagnosis" field (Figure 3A), the application reached 88.7% of accuracy, recall and F1-score, while achieving 100% of precision. For the "biopsy date" (Figure 3B) and the "ID number" (of the report, Figure 3C) fields all the applied metrics were >90%. High performance was obtained for the "specimen type" field (87.0% of accuracy and precision, 87.3% of F1-score and 98.5% of precision, Figure 3D). High performance levels were achieved for "IHC execution" field (94.6% of accuracy, recall and F score, and 100% of precision, Figure 3E), although focusing the analysis on single biomarker, accuracy, recall and F1-score decreased, while retaining nearly 100% of precision (Supplementary Table S3). Similar results were noticed also for data concerning "BM and FISH execution" (Figure 3F and Figure 3G). Finally, the tool allowed the detection of Ki-67-related information with 82.6%, 99.4%, 78.9% and 80.7% of accuracy, precision, recall and F1-score, respectively (Figure 3H).
In order to demonstrate OCR limitation in detecting each single biomarker, we selected first fifty reports according to image resolution and we reassessed validation metrics (Table 1). Overall, comparing performances with those assessed from all reports, recall and F1-score remarked by an averaged improvement of 17.2% and 12.4%, respectively.

Table 1. ARGO performance between all reports and the top 50 reports[a].

	PRECISION (%)			RECALL (%)			F1 SCORE (%)		
	All reports	Best reports*	Diff	All reports	Best reports*	Diff	All reports	Best reports*	Diff
DATA-FIELD									
MYC	100.0	100.0	0	48.6	69.2	20.6	63.7	78.9	15.2
BCL2	98.5	100.0	1.5	69.6	84.8	15.2	72.5	87.3	14.8
BCL6	98.3	100.0	1.7	61.1	84.4	23.2	64.4	87.1	22.7
CD10	97.0	100.0	3.0	55.7	78.1	22.4	60.5	81.9	21.4
CD20	99.4	100.0	0.6	74.4	92.3	17.9	75.5	93.1	17.6
Cyclin D1	100.0	100.0	0	58.8	62.5	3.7	80.5	75.1	-5.4
	Mean (std)		1.4 (1.1)	Mean (std)		17.2 (7.2)	Mean (std)		12.4 (10.2)

[a]Top 50 reports with high optical resolution.

Figure 1. Graphical description of the framework.
A) Each paper-based report is manually transformed into an image file by a common digital scanner (right upside, an example of paper-based report from the Pathology Unit of the IRCCS "Giovanni Paolo II" of Bari, Italy). Then, the image is uploaded into the ARGO web application (black block), transformed in structured text through OCR and saved (by a NLP approach) as structured data in a database via webserver. "Diagnosis" attribution is carried out via API connecting ARGO with SEER servers (blue block). Finally, ARGO automatically populates eCRFs via API (red block). **B)** Representative picture of REDCap dashboard for a single case report including "Demography" and "Disease parameters" forms (red bullets).

Figure 2. Characteristics of data retrieved from diagnostic reports.
Graphical representation of diagnostic features, subdivided into specific fields, captured by ARGO from a total of 239 paper-based pathology reports of DLBCL, FL, and MCL.

Figure 3. ARGO performance.
Series of radar graphs indicating the performance metric as percentage of accuracy, precision, recall and F1 score for different fields.

Discussion

Principal Results

We aimed this study at designing a workflow to automate the collection of RW onco-hematological data, with particular regard to lymphoma diagnoses. We developed an NLP-based application, called ARGO, and provided a “proof of concept” for its reliability in generating eCRFs directly from unstructured histopathology reports. We successfully tested ARGO performance, in terms of accuracy, precision, recall and F1 score, on a monocentric cohort of 239 lymphoma cases including DLBCL, FL, and MCL.

In comparison with other applications in oncology, ARGO confirmed super-imposable performances in data field detection (6,11,14,16,30,31), while overcoming some limitations. For instance, in the work by Nguyen et al., each metric decreases as the number of classes describing a certain data field increases (11). This trend is globally confirmed in our experience, and even for data fields with high number of classes, such as “Specimen Type”, we achieved very high precision level. To potentiate the OCR performance, we created three separate thesauri for “biomarkers”, “specimens” and “diagnosis”. As in Tanenblatt et al. (31), we first included officially-recognized nomenclatures in the “biomarkers” and “diagnosis” dictionaries, referring to the ICD10 classification. Then, we manually added synonyms, abbreviations and other uncommon expressions noticed in our set of reports. Nevertheless, ARGO failed in converting six reports as a direct consequence of OCR-based limitations in reading reports with low-quality optical resolution.

From a more applicative point of view, ARGO might maximize the use of clinical data in translational research by boosting the adoption of EHR. Especially in onco-hematology, the public healthcare system still lacks standardized models of RW data collection, and a number of gaps exist concerning how to electronically collect unstructured information. Application of a computerized approach to extract data from paper-based reports and directly populate eCRFs provides two main advantages, i) standardization of data collection; and ii) data integration between Institutions and research networks. Finally, our system takes advantage from two levels of personalization related to REDCap, i) the designing of graphic interfaces directly by the clinical investigators according to specific endpoints; and ii) the easily population of eCRFs via API. Therefore, ARGO appeared as a valid tool for a precise and time-saving recording of clinical data when compared to manual abstraction (16). Our approach results feasible in the daily practice, facilitating consultation, filtering and management of RW data. This step is crucial to study wide proportions of onco-hematological patients who have no access to clinical trials and support national research networking.

Limitations

Main limitations of the study could be the monocentric source of the histopathology report template and the language. However, current pathology reporting systems allow the use of personalized data fields according to shared templates and translational software, e.i. as “MyMemory” software, enable the easy switch up across languages.

Conclusions

In conclusion, although currently limited to a monocentric subset of lymphoma subtypes, our approach could be tailored to additional disease models in oncology and might also be feasible for future machine-learning applications.

Acknowledgements

The Authors are grateful to Eng. Giancarlo Salomone and Francesco Pacoda for their technical help

as members of the ICT staff of IRCCS Istituto “Giovanni Paolo II”. The study was funded by Italian Minister of Health (Grant RC2018-2020) and the Apulia Region Grant “TecnoMED, Tecnopolo della medicina di precisione”.

Conflicts of Interest

none declared.

Abbreviations

ARGO: Automatic Record Generator for Oncology
RW: Real-world
NLP: natural language processing
OCR: optical character recognition
EHR: Electronic health records
eCRFs: electronic case report forms
DLBCL: diffuse large b-cell lymphoma
FL: follicular lymphoma
MCL: mantle cell lymphoma
NHL: non-Hodgkin lymphoma
ID: identification code
IHC: immunohistochemical
LN: lymph-node
EN: extra-nodal
BM: bone marrow
PB: peripheral blood
CD: cluster of differentiation
FISH: fluorescent *in situ* hybridization
COO: cell of origin
SEER: surveillance, epidemiology, and end result
ICD10: international statistical classification of disease and related health problems 10^{th} version
API: application programming interface
Diff: difference
Std: standard deviation
GCB: germinal center B-like

Multimedia Appendix 1

Supplementary_appendix.pdf

Multimedia Appendix 2

ARGO_demo_final.mp4

References

1. Swerdlow SH, Campo E, Harris NL, Jaffe ES, Pileri SA, Stein H, et al., editors. WHO Classification of Tumours of Haematopoietic and Lymphoid Tissues [Internet]. Revised 4t. Lyon (FR): IARC Press; 2017. Available from: https://publications.iarc.fr/Book-And-Report-

Series/Who-Classification-Of-Tumours/WHO-Classification-Of-Tumours-Of-Haematopoietic-And-Lymphoid-Tissues-2017
2. Campo E, Swerdlow SH, Harris NL, Pileri S, Stein H, Jaffe ES. The 2008 WHO classification of lymphoid neoplasms and beyond: Evolving concepts and practical applications. Blood [Internet]. 2011;117(19):5019–32. Available from: https://doi.org/10.1182/blood-2011-01-293050%0A
3. Khozin S, Blumenthal GM, Pazdur R. Real-world Data for Clinical Evidence Generation in Oncology. J Natl Cancer Inst [Internet]. 2017;109(11):1–5. Available from: https://doi.org/10.1093/jnci/djx187
4. Zaccaria GM, Ferrero S, Rosati S, Ghislieri M, Genuardi E, Evangelista A, et al. Applying data warehousing to a phase III clinical trial from the Fondazione Italiana Linfomi (FIL) ensures superior data quality and improved assessment of clinical outcomes. JCO Clin Cancer Informatics [Internet]. 2019;1–15. Available from: https://ascopubs.org/doi/full/10.1200/CCI.19.00049
5. Zong N, Wen A, Stone DJ, Sharma DK, Wang C, Yu Y, et al. Developing an FHIR-Based Computational Pipeline for Automatic Population of Case Report Forms for Colorectal Cancer Clinical Trials Using Electronic Health Records. JCO Clin Cancer Informatics [Internet]. 2020;4(4):201–9. Available from: https://ascopubs.org/doi/pdf/10.1200/CCI.19.00116
6. Xie F, Chen Q, Zhou Y, Chen W, Bautista J, Nguyen ET, et al. Characterization of patients with advanced chronic pancreatitis using natural language processing of radiology reports. PLoS One [Internet]. 2020;15(8 August):1–13. Available from: http://dx.doi.org/10.1371/journal.pone.0236817
7. Buckley JM, Coopey SB, Sharko J, Polubriaginof F, Drohan B, Belli AK, et al. The feasibility of using natural language processing to extract clinical information from breast pathology reports. J Pathol Inform [Internet]. 2012/06/30. 2012;3:23. Available from: https://pubmed.ncbi.nlm.nih.gov/22934236
8. Zhang Y, Cai T, Yu S, Cho K, Hong C, Sun J, et al. High-throughput phenotyping with electronic medical record data using a common semi-supervised approach (PheCAP). Nat Protoc [Internet]. 2019;14(12):3426–44. Available from: http://dx.doi.org/10.1038/s41596-019-0227-6
9. Venkataraman GR, Pineda AL, Bear Don't Walk OJ, Zehnder AM, Ayyar S, Page RL, et al. FasTag: Automatic text classification of unstructured medical narratives. PLoS One [Internet]. 2020;15(6 June):1–18. Available from: http://dx.doi.org/10.1371/journal.pone.0234647
10. Ryu B, Yoon E, Kim S, Lee S, Baek H, Yi S, et al. Transformation of Pathology Reports into the Common Data Model with Oncology Module: Use Case for Colon Cancer (Preprint). J Med Internet Res [Internet]. 2020;22. Available from: https://www.jmir.org/2020/12/e18526
11. Nguyen AN, Moore J, O'Dwyer J, Philpot S. Assessing the Utility of Automatic Cancer Registry Notifications Data Extraction from Free-Text Pathology Reports. AMIA . Annu Symp proceedings AMIA Symp. 2015;2015:953–62.
12. Spasić I, Livsey J, Keane JA, Nenadić G. Text mining of cancer-related information: Review of current status and future directions. Int J Med Inform [Internet]. 2014;83(9):605–23. Available from: https://www.sciencedirect.com/science/article/pii/S1386505614001105?via%3Dihub
13. Lin FPY, Pokorny A, Teng C, Epstein RJ. TEPAPA: A novel in silico feature learning pipeline for mining prognostic and associative factors from text-based electronic medical records. Sci Rep [Internet]. 2017;7(1):1–13. Available from: http://dx.doi.org/10.1038/s41598-017-07111-0
14. Glaser AP, Jordan BJ, Cohen J, Desai A, Silberman P, Meeks JJ. Automated Extraction of Grade, Stage, and Quality Information From Transurethral Resection of Bladder Tumor Pathology Reports Using Natural Language Processing. JCO Clin Cancer Informatics [Internet]. 2018;(2):1–6. Available from: https://ascopubs.org/doi/10.1200/CCI.17.00128

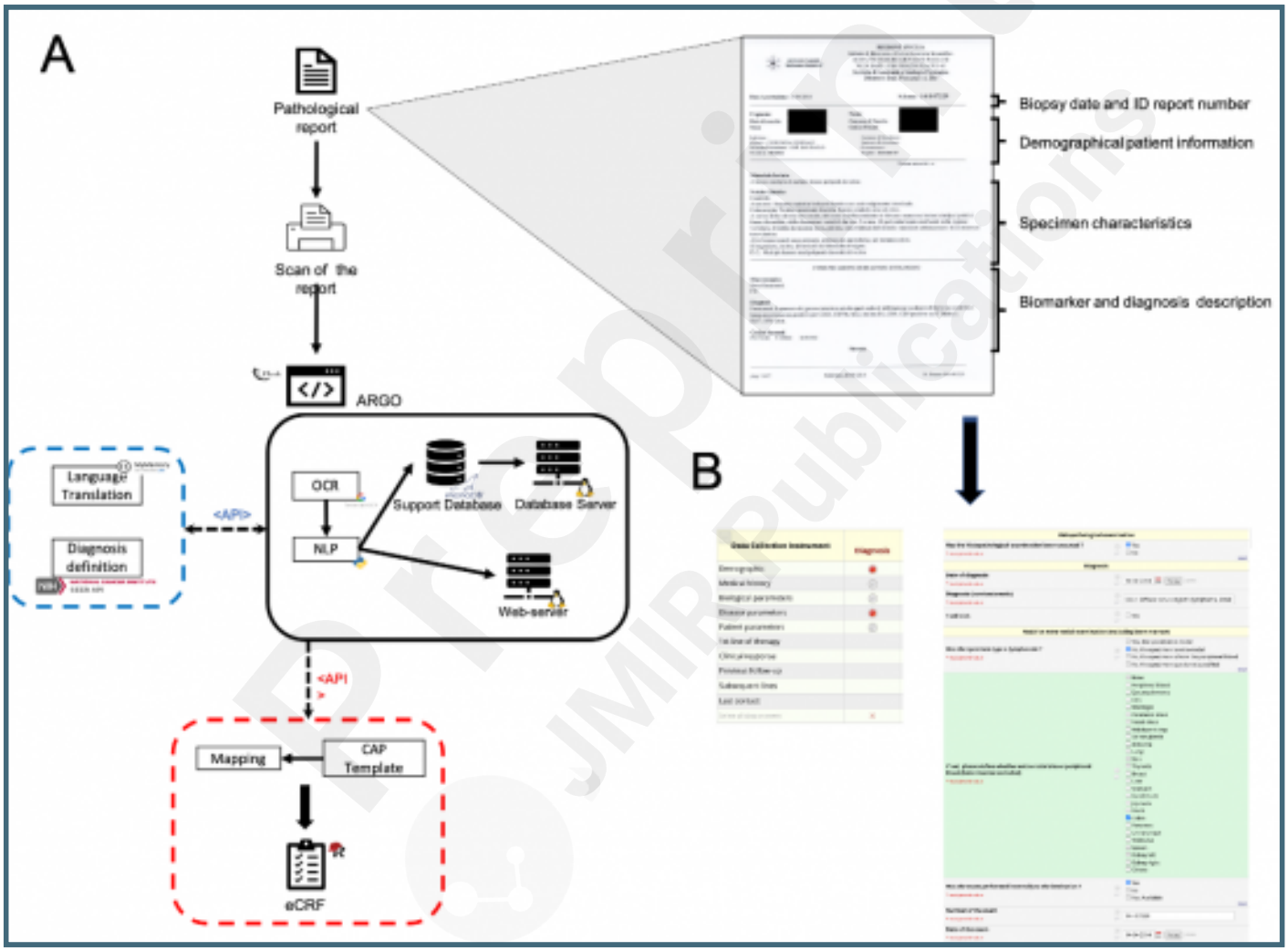
15. Odisho AY, Park B, Altieri N, DeNero J, Cooperberg MR, Carroll PR, et al. Natural language processing systems for pathology parsing in limited data environments with uncertainty estimation. JAMIA Open [Internet]. 2020;3(3):431–8. Available from: https://doi.org/10.1093/jamiaopen/ooaa029
16. Odisho AY, Bridge M, Webb M, Ameli N, Eapen RS, Stauf F, et al. Automating the Capture of Structured Pathology Data for Prostate Cancer Clinical Care and Research. JCO Clin Cancer Informatics [Internet]. 2019;(3):1–8. Available from: https://ascopubs.org/doi/10.1200/CCI.18.00084
17. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)-A metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform [Internet]. 2009;42(2):377–81. Available from: http://dx.doi.org/10.1016/j.jbi.2008.08.010
18. Harris PA, Taylor R, Minor BL, Elliott V, Fernandez M, O'Neal L, et al. The REDCap consortium: Building an international community of software platform partners. J Biomed Inform [Internet]. 2019;95(April):103208. Available from: https://doi.org/10.1016/j.jbi.2019.103208
19. Moon S, Liu S, Kingsbury P, Chen D, Wang Y, Shen F, et al. Medical concept intersection between outside medical records and consultant notes: A case study in transferred cardiovascular patients. Proc - 2017 IEEE Int Conf Bioinforma Biomed BIBM 2017 [Internet]. 2017;2017-Janua(November):1495–500. Available from: https://ieeexplore.ieee.org/document/8217883
20. Qader WA, Ameen MM. Diagnosis of Diseases from Medical Check-up Test Reports Using OCR Technology with BoW and AdaBoost algorithms. Proc 5th Int Eng Conf IEC 2019 [Internet]. 2019;(June):205–10. Available from: https://ieeexplore.ieee.org/document/8950605
21. Hans CP, Weisenburger DD, Greiner TC, Gascoyne RD, Delabie J, Ott G, et al. Confirmation of the molecular classification of diffuse large B-cell lymphoma by immunohistochemistry using a tissue microarray. Blood [Internet]. 2004;103(1):275–82. Available from: https://doi.org/10.1182/blood-2003-05-1545%0A
22. Schantz HF. The history of OCR, optical character recognition. [Manchester Center, Vt.]: Recognition Technologies Users Association; 1982.
23. Eisenstein J. Introduction to Natural Language Processing. Adaptive Computation and Machine Learning serie. MIT Press; 2019. 536 p.
24. Open Source Computer Vision Library - Miscellaneous Image Transformations [Internet]. Available from: https://opencv.org/
25. Ruhl J, Adamo MP, Dickie L, Negoita S. Hematopoietic and Lymphoid Neoplasm Coding Manual [Internet]. Bethesda, MD, US; 2020. Available from: https://seer.cancer.gov/tools/heme/Hematopoietic_Instructions_and_Rules.pdf
26. World Health Organization. Classification of diseases (ICD) [Internet]. Available from: https://www.who.int/classifications/classification-of-diseases
27. Duncavage E, Advani RH, Agosti S, Foulis P, Gibson C, Kang L, et al. Template for reporting results of biomarker testing of specimens from patients with diffuse large B-cell lymphoma, not otherwise specified. Arch Pathol Lab Med. 2016;140(11):1225–7.
28. Ellis DW. Protocol for the Examination of Specimens From Patients With Hodgkin Lymphoma. 2013;(October):1–15.
29. Resnik P, Lin J. Evaluation of NLP Systems. In: Clark A, Fox C, Lappin S, editors. The Handbook of Computational Linguistics and Natural Language Processing. Wiley-Blackwell; 2010. p. 271–95.
30. Hanauer DA, Miela G, Chinnaiyan AM, Chang AE, Blayney DW. The Registry Case Finding Engine: An Automated Tool to Identify Cancer Cases from Unstructured, Free-Text Pathology Reports and Clinical Notes. J Am Coll Surg [Internet]. 2007;205(5):690–7. Available from:

https://www.sciencedirect.com/science/article/pii/S1072751507006382?via%3Dihub
31. Tanenblatt M, Coden A, Sominsky I. The ConceptMapper approach to named entity recognition. Proc 7th Int Conf Lang Resour Eval Lr 2010. 2010;546–51.

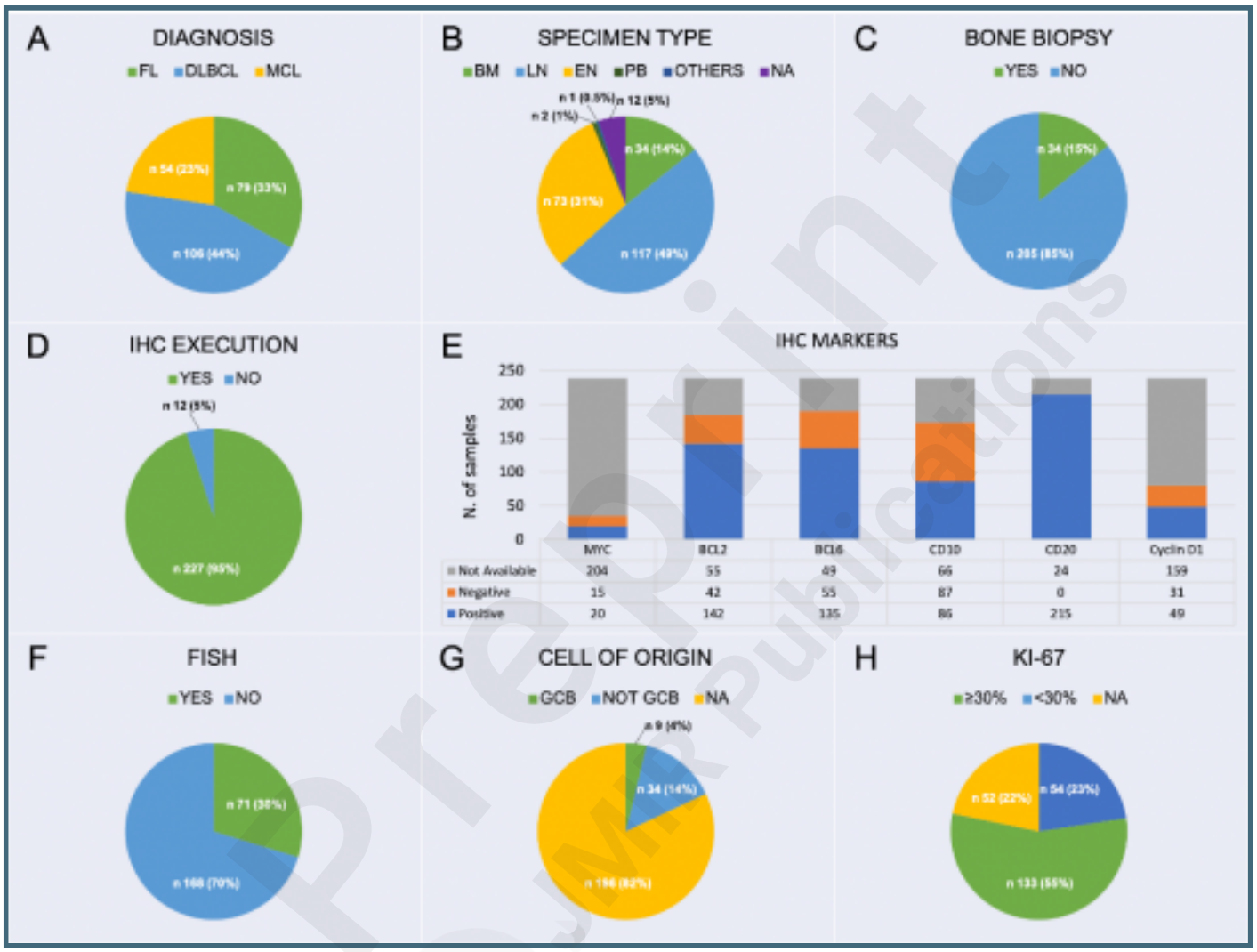
Supplementary Files

Figures

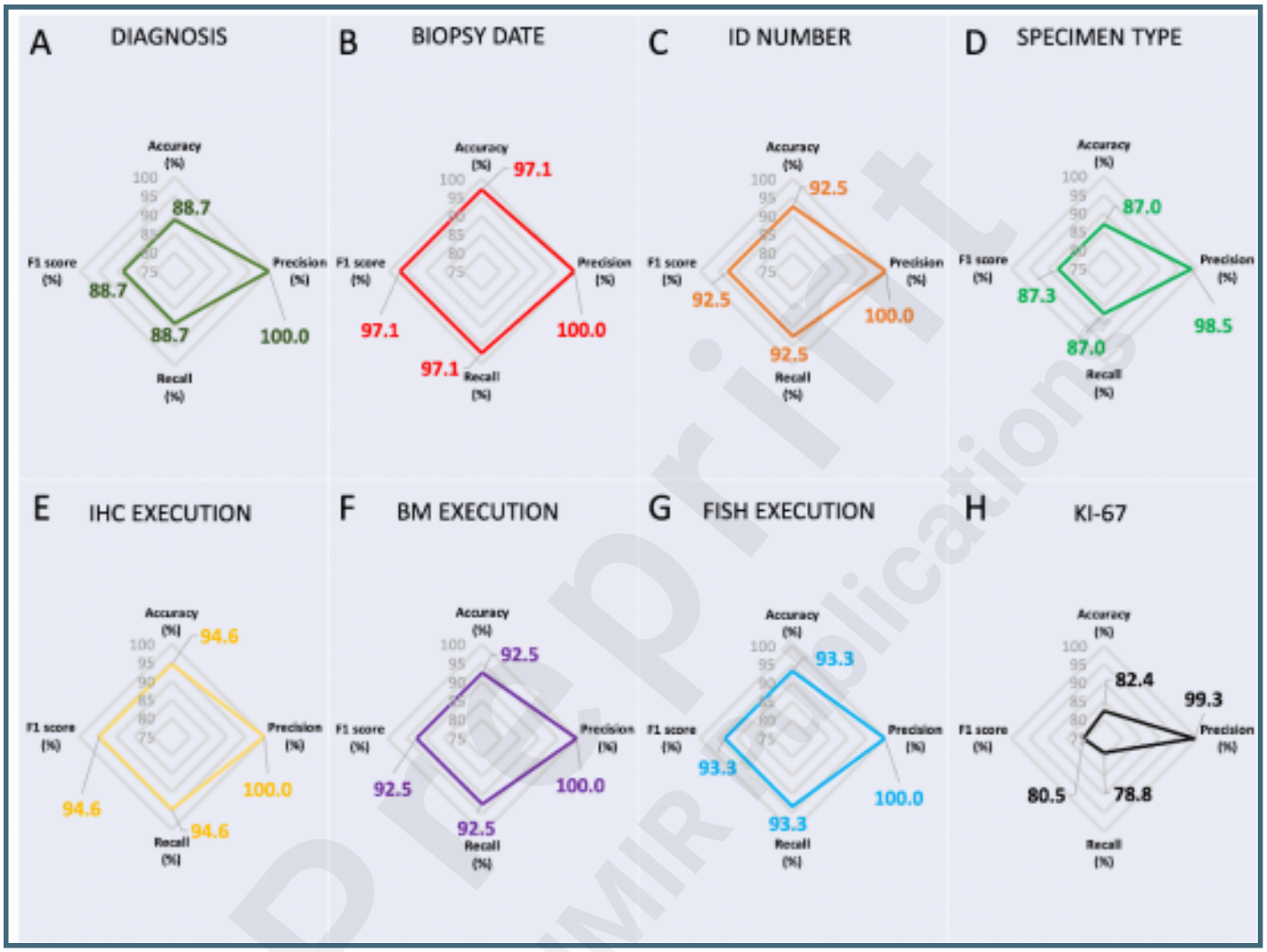
Graphical description of the framework A) Each paper-based report is manually transformed into an image file by a common digital scanner (right upside, an example of paper-based report from the Pathology Unit of the IRCCS “Giovanni Paolo II” of Bari, Italy). Then, the image is uploaded into the ARGO web application (black block), transformed in structured text through OCR and saved (by a NLP approach) as structured data in a database via webserver. “Diagnosis” attribution is carried out via API connecting ARGO with SEER servers (blue block). Finally, ARGO automatically populates eCRFs via API (red block). B) Representative picture of REDCap dashboard for a single case report including “Demography” (red bullets) and “Disease parameters” forms. Abbreviations. ARGO, automatic record generator for oncology; OCR, optical character recognition; NLP, natural language process; API, application programming interface; SEER, Surveillance, Epidemiology, and End Result; NIH, national institute of health; CAP, college of American pathologists; eCRF, electronic case report form.



Characteristics of data retrieved from diagnostic reports. Graphical representation of diagnostic features, subdivided into specific fields, captured by ARGO from a total of 239 paper-based pathology reports of DLBCL, FL, and MCL. Abbreviations. FL, follicular lymphoma; DLBCL, diffuse large B-cell lymphoma; MCL, Mantle cell lymphoma; BM, bone marrow; LN, lymph-node; EN, extra nodal; PB, peripheral blood; BM, bone marrow; NA: not available; CD, cluster of differentiation; FISH, fluorescent in situ hybridization; GCB, germinal center B-like.



ARGO performance. Series of radar graphs indicating the performance metric as percentage of accuracy, precision, recall and F1 score for different fields. Abbreviations. N, number; IHC, immunohistochemical; BM, bone marrow; CD, cluster of differentiation; FISH, fluorescent in situ hybridization.



Multimedia Appendixes

Supplementary appendix.
URL: https://asset.jmir.pub/assets/371c78d24121457c9f492ce940ae6067.pdf

ARGO demonstration.
URL: https://asset.jmir.pub/assets/9bd6dba387a3f18cab88b5db44f0053a.mp4