ELSEVIER

Blood 142 (2023) 122–124

The 65th ASH Annual Meeting Abstracts

ORAL ABSTRACTS

803.EMERGING TOOLS, TECHNIQUES AND ARTIFICIAL INTELLIGENCE IN HEMATOLOGY

Clinical Text Reports to Stratify Patients Affected with Myeloid Neoplasms Using Natural Language Processing

Gianluca Asti, MSc[1], Elisabetta Sauta, PhD[1], Nico Curti, PhD[2,3], Gianluca Carlini, PhD[2,3], Lorenzo Dall'Olio, PhD[3,2], Luca Lanino, MD[4], Giulia Maggioni, MD[1], Alessia Campagna, MD[1], Marta Ubezio, MD[1], Antonio Russo, MD[1], Gabriele Todisco, MD[1], Cristina Astrid Tentori, MD[1], Pierandrea Morandini, MEng[1], Marilena Bicchieri, PhD[1], Maria Chiara Grondelli, BSc[1], Matteo Zampini, PhD[1], Erica Travaglino, BS[1], Victor Savevski, MEng[1], Nicolas Riccardo Derus, PhD[5], Daniele Dall'Olio, PhD[5], Claudia Sala, PhD[5], Lin-Pierre Zhao, MD[6], Armando Santoro, MD[1], Shahram Kordasti, MDPhD[7], Valeria Santini, MD[8], Anne Sophie Kubasch, MD[9], Uwe Platzbecker, MD[10], Maria Diez-Campelo, MD PhD[11], Pierre Fenaux, MD PhD[6], Amer M. Zeidan, MBBS, MHS[12], Torsten Haferlach, MD PhD[13], Gastone Castellani, PhD[5], Matteo Giovanni Della Porta, MD[1], Saverio D'Amico, MSc[14]

[1] Humanitas Clinical and Research Center, IRCCS, Rozzano, Italy
[2] Department of Physics and Astronomy, University of Bologna, Bologna, Italy
[3] Data Science and Bioinformatics Laboratory, IRCCS Institute of Neurological Sciences of Bologna, Bologna, Italy
[4] Humanitas Clinical and Research Center, IRCCS, Rozzano, Italy
[5] University of Bologna, Bologna, Italy
[6] Saint Louis hospital APHP, Paris, France
[7] King's College London, London, United Kingdom
[8] MDS Unit, DMSC, AOU Careggi, University of Florence, Firenze, Italy
[9] Department of Hematology, Cellular Therapy, Hemostaseology and Infectious Diseases, University Medical Center Leipzig, Leipzig, Germany
[10] Department of Hematology, Cellular Therapy, Hemostaseology and Infectious Diseases, University Leipzig Medical Center, Leipzig, Germany
[11] Department of Hematology, Salamanca-IBSAL University Hospital, Salamanca, Spain
[12] Section of Hematology, Department of Internal Medicine, Yale School of Medicine and Yale Cancer Center, New Haven, CT
[13] MLL Munich Leukemia Laboratory, Munich, Germany
[14] Humanitas Clinical and Research Center, IRCCS, Rozzano (Milan), Italy

Background: The availability of multimodal patient data, such as demographics, clinical, imaging, treatment, quality of life, outcomes and wearables data, as well as genome sequencing, have paved the way for the development of multimodal clinical solutions that introduce personalized or precision medicine. The clinical report is an information layer that contains relevant information about the disease in addition to the patient's point of view. Natural language processing (NLP) is a branch of artificial intelligence (AI) and its pre-trained language models are the key technology for extracting value from this data layer.
Aims: This project was conducted by GenoMed4all and Synthema EU consortia, with the aim to: 1) Build an AI language model specific for the hematology domain. 2) Use NLP technology to extract relevant information from clinical reports and perform unsupervised stratification of patients, in order to 3) demonstrate that the clinical report is earlier access to data relative to disease clinical phenotype and biology and provide important information for patient stratification and prediction of clinical outcomes.
Methods: To translate text sentences into numerical embeddings, we implemented bidirectional encoder representations from transformers (BERT) framework. To learn text representations and correlations within data, we performed domain-adaptation by fine-tuned pre-trained model on hematological clinical reports of patients with myeloproliferative neoplasms (MPN), myelodysplastic syndrome (MDS) and acute myeloid leukemia (AML). Patient stratification was performed by HDB-SCAN clustering on text embedding encoded by BERT (HematoBERT). Clusters validation was performed by assessing patients' diagnosis and survival probability. Finally, we compared domain-tuned HematoBERT vs pre-trained non-contextualized models.

Results: We implemented HematoBERT based on the bert-base-multilingual-uncased version of BERT. Training data were hematological text reports of 1,328 patients. During fine-tuning, texts were tokenized, then we randomly replaced 15% of the tokens with masked tokens, training the model to predict them. We performed stratification using clinical reports from a validation cohort of 360 patients. We identified 7 clusters, defined according to similar words in meaning that were placed in a specific topic. We extracted the most important words and concepts for each cluster (topic) and we summarized them into effective descriptions for each group of patients. Two clusters included MDS patients with excess blasts, and without excess blasts with ring sideroblasts and del5q (n=69, n=115). One cluster included patients with excess blasts and MDS/MPN (n= 33). Two clusters included MPN patients with primary and secondary myelofibrosis, and MPN patients most including subjects affected with polycythemia vera and essential thrombocythemia (n=35, n=46). Two clusters included patients with AML from MDS and therapy-related AML, and patients with de novo AML (n=22, n=42). Clinical validation was performed based on the diagnosis and survival probability of patients assigned to clusters. Patients' diagnoses were compatible with the cluster assignment (Figure 1). Frequency of gene mutations (as assessed by targeted Next-Generation Sequencing) among different clusters reflected the well-known genotypic-phenotypic associations in MDS, MPN and AML. Kaplan-Maier curves indicated significative risk stratification in clusters in terms of survival probability (Figure 2), similar to stratifications performed on clinical and genomic data. Finally, we evaluate the domain adaptation by comparing the model to other pre-trained non-contextualized ones. Pseudo perplexity score (PPS), accuracy and F1 score were calculated to quantify how good the models are when they see new data, predicting the next word given the context of the sentence. HematoBERT obtained high PPS, accuracy and F1 scores, outperforming the other models also trained on generic clinical domains.
Conclusion: Domain-adapted language models are able to understand contexts and correlations in documents. HematoBERT can be used to extract relevant features from clinical reports. This data layer is relevant to perform disease stratification of patients based on clinical and genomic information and could be integrated into next-generation multimodal models of personalized medicine.

Disclosures Santoro: *Amgen:* Speakers Bureau; *Abbvie:* Speakers Bureau; *Roche:* Speakers Bureau; *Takeda:* Speakers Bureau; *Merck MSD:* Membership on an entity's Board of Directors or advisory committees, Speakers Bureau; *Bayer:* Membership on an entity's Board of Directors or advisory committees, Speakers Bureau; *Eisai:* Membership on an entity's Board of Directors or advisory committees, Speakers Bureau; *Pfizer:* Membership on an entity's Board of Directors or advisory committees, Speakers Bureau; *Gilead:* Membership on an entity's Board of Directors or advisory committees, Speakers Bureau; *Servier:* Membership on an entity's Board of Directors or advisory committees, Speakers Bureau; *BMS:* Membership on an entity's Board of Directors or advisory committees, Speakers Bureau; *Incyte:* Consultancy; *Sanofi:* Consultancy; *Celgene (BMS):* Speakers Bureau; *AstraZeneca:* Speakers Bureau; *Eli Lilly:* Speakers Bureau; *Sandoz:* Speakers Bureau; *Novartis:* Speakers Bureau; *Arqule:* Other. **Kordasti:** *Novartis:* Honoraria, Membership on an entity's Board of Directors or advisory committees; *Beckman Coulter:* Honoraria; *MorphoSys:* Research Funding. **Santini:** *BMS, Abbvie, Geron, Gilead, CTI, Otsuka, servier, janssen, Syros:* Membership on an entity's Board of Directors or advisory committees. **Platzbecker:** *Servier:* Consultancy, Honoraria, Research Funding; *Geron:* Consultancy, Research Funding; *MDS Foundation:* Membership on an entity's Board of Directors or advisory committees; *Celgene:* Honoraria; *Merck:* Research Funding; *Syros:* Consultancy, Honoraria, Research Funding; *Fibrogen:* Research Funding; *Amgen:* Consultancy, Research Funding; *Novartis:* Consultancy, Honoraria, Research Funding; *Jazz:* Consultancy, Honoraria, Research Funding; *Takeda:* Consultancy, Honoraria, Research Funding; *AbbVie:* Consultancy; *Curis:* Consultancy, Research Funding; *Silence Therapeutics:* Consultancy, Honoraria, Research Funding; *Janssen Biotech:* Consultancy, Research Funding; *BMS:* Research Funding; *Bristol Myers Squibb:* Consultancy, Honoraria, Membership on an entity's Board of Directors or advisory committees, Other: travel support; medical writing support, Research Funding; *BeiGene:* Research Funding; *Roche:* Research Funding. **Diez-Campelo:** *BMS/Celgene:* Consultancy, Honoraria, Membership on an entity's Board of Directors or advisory committees, Other: Advisory board fees; *Novartis:* Consultancy, Honoraria, Membership on an entity's Board of Directors or advisory committees; *Gilead Sciences:* Other: Travel expense reimbursement; *GSK:* Consultancy, Membership on an entity's Board of Directors or advisory committees. **Fenaux:** *Janssen:* Consultancy, Honoraria, Research Funding; *AbbVie:* Consultancy, Honoraria, Research Funding; *Bristol Myers Squibb:* Consultancy, Honoraria, Research Funding; *Novartis:* Consultancy, Honoraria, Research Funding; *Jazz:* Consultancy, Honoraria, Research Funding; *French MDS Group:* Honoraria. **Zeidan:** *Shattuck Labs:* Research Funding; *Gilead:* Consultancy, Honoraria; *Celgene/BMS:* Consultancy, Honoraria; *AbbVie:* Consultancy, Honoraria; *Astex:* Research Funding; *Incyte:* Consultancy, Honoraria; *Lox Oncology:* Consultancy, Honoraria; *Foran:* Consultancy, Research Funding; *BeyondSpring:* Consultancy, Honoraria; *BioCryst:* Consultancy, Honoraria; *Notable:* Consultancy, Honoraria; *Kura:* Consultancy, Honoraria; *Tyme:* Consultancy, Honoraria; *Schrödinger:* Consultancy, Honoraria; *Zentalis:* Consultancy, Honoraria; *Mendus:* Consultancy, Honoraria; *Orum:* Consultancy, Honoraria; *Syndax:* Consultancy, Honoraria; *Epizyme:* Consultancy, Honoraria; *Genentech:* Consultancy, Honoraria; *Janssen:* Consultancy, Honoraria; *Amgen:* Consultancy, Honoraria; *Taiho:* Consultancy, Honoraria; *Geron:* Consultancy, Honoraria; *Daiichi Sankyo:* Consultancy, Honoraria; *Astellas:* Consultancy, Honoraria; *Novartis:* Consultancy, Honoraria; *Boehringer-Ingelheim:* Consultancy, Honoraria; *Servier:* Consultancy, Honoraria; *Agios:* Consultancy, Honoraria; *Pfizer:* Consultancy, Honoraria; *Seattle Genetics:* Consultancy, Honoraria; *Ionis:* Consultancy, Honoraria; *Takeda:* Consultancy, Honoraria; *Otsuka:* Consultancy, Honoraria; *Chiesi:* Consultancy, Honoraria; *ALX Oncology:* Consultancy, Honoraria; *Regeneron:* Consultancy, Honoraria; *Jazz:* Consultancy, Honoraria; *Syros:* Consultancy, Honoraria. **Haferlach:** *MLL Munich Leukemia Laboratory:* Current Employment, Other: Equity Ownership. **Della Porta:** *Bristol Myers Squibb:* Honoraria, Membership on an entity's Board of Directors or advisory committees.

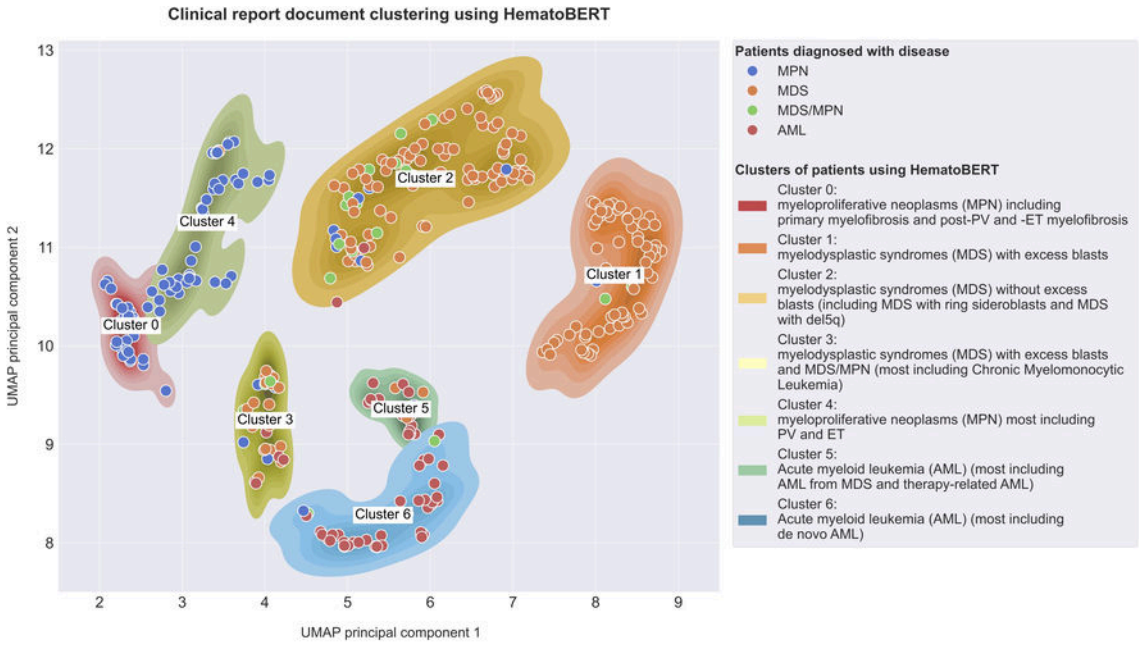

Figure 1. Clinical reports-based clustering of patients with MDS, MPN and AML using HematoBERT. UMAP 2-dimensional text embedding encoded by fine-tuned HematoBERT (each dot represents a patient) whose location is defined on the basis of the textual representation and correlation. The figure shows the clustering of patients' (noise cluster was excluded) clinical records according to similar words and concepts. Stratification of patients by diagnosis is consistent with cluster groupings.

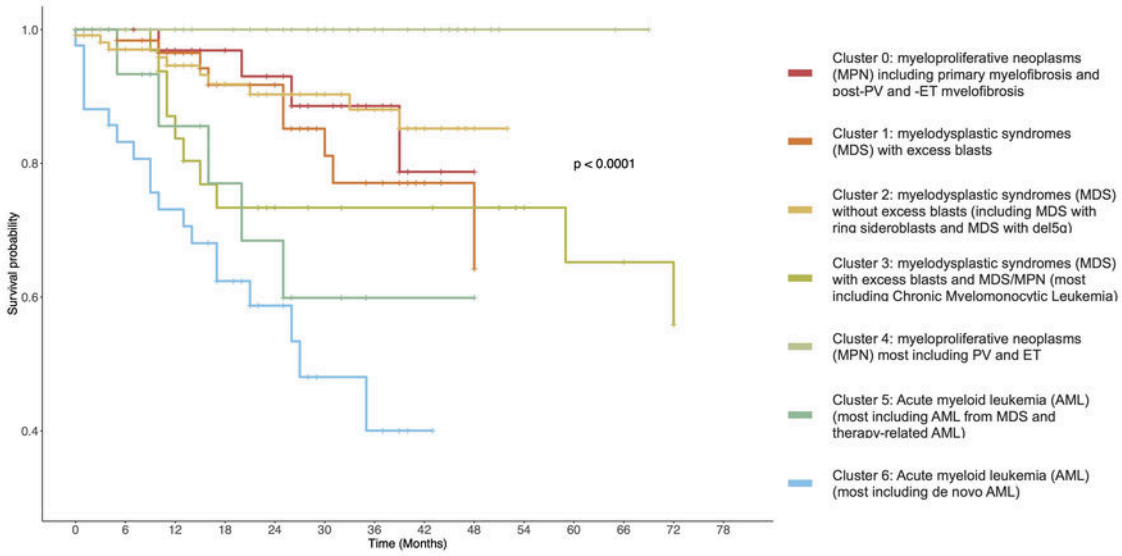

Figure 2. Kaplan-Meier probability estimates of Overall Survival (OS) stratified by clusters identified using HematoBERT.

Figure 1

https://doi.org/10.1182/blood-2023-188292