RWD55

Digital poster

Deep Real-World Analysis of Patients With Myelofibrosis Using Natural Language Processing and Machine Learning: A Methods Description

Fulya Sen Nikitas,[1] Tim d'Estrubé,[1] Bella Vo,[2] Paul Juneau,[3] Julien Graff,[4] Shiyuan Zhang,[2] Margarita Posso,[5] María Lopez,[5] Marco Garranzo,[5] Lucia Cabal-Hierro,[5] Berta Biescas[5]

[1]GSK, London, UK; [2]GSK plc, Philadelphia, PA, USA; [3]GSK, Boyds, MD, USA; [4]GSK, Baar, Zug, Switzerland; [5]Savana Research, Madrid, Spain

Background

- Myelofibrosis (MF) is a myeloproliferative neoplasm characterized by the pathological proliferation of stem cell–derived myeloid cells[1,2] with various clinical manifestations, including severe anemia and constitutional symptoms
- The clinically heterogeneous manifestations of this disease and lack of real-world data (RWD) underscore the need for techniques that can help supplement existing trial data to inform clinical decision-making in MF
- This study employs clinical natural language processing (cNLP) techniques to extract and analyze RWD from electronic health records (EHRs)
- The study objectives include the following:
  - Characterization of patients with MF and their treatment patterns across a wide range of clinical settings
  - Analysis of comorbidities, clinical characteristics, healthcare resource utilization, clinical outcomes, and factors linked to disease progression and mortality
- This protocol will serve as an umbrella framework for upcoming studies

Methods

Study Design

- This multicenter, observational cohort study is a retrospective analysis of EHRs from hospitals in Spain, the UK, Austria, and France (2015-2027)
- The data sources include structured and unstructured clinical data extracted from EHRs

cNLP System and MF Disease Panel

EHRead technology features

- A dedicated cNLP pipeline trained on real-world clinical text
- Used for extraction of >230 clinical concepts organized into a terminology-based MF Disease Panel
- The panel is a predefined list of clinically relevant terms aligned with the study objectives
- The system leverages 3 global medical terminologies to ensure consistent representation of diagnoses, treatments, procedures, laboratory values, and medications

Standardized terminologies

SNOMED CT
Systematized Nomenclature of Medicine – Clinical Terms

ATC
Anatomical Therapeutic Chemical classification system

LOINC
Logical Observation Identifiers Names and Codes

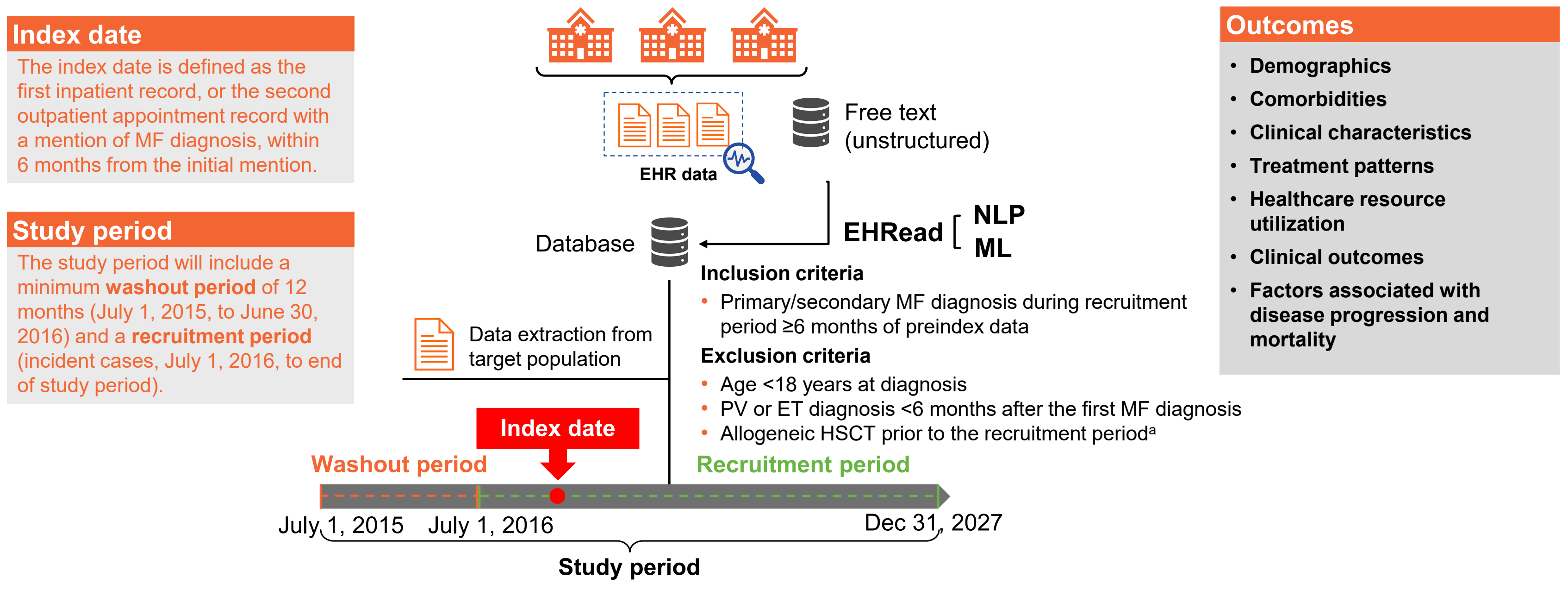
Data Extraction and Analysis

- Named-Entity Linking (NEL) models were used to identify clinical terms
- Specialized cNLP models were used for robust data extraction (ie, negation, temporality, relations)

Validation Process and Performance Evaluation

- Validation of cNLP extraction was conducted by medical annotators and external experts at participating centers
- cNLP performance was evaluated on precision, recall, and F1 score free-text samples selected via SLiCE (Smart Linguistic Corpus Extraction), a proprietary tool for robust sampling
  - The F1 score was calculated as (2 × Precision × Recall)/(Precision + Recall) and is the overall performance indicator of information retrieval. The F1 score ranges between 0 and 1, where 1 represents a perfect model and 0 represents a complete failure
  - Reannotation cycles were triggered when the F1 score was <0.8 on PIO terms
- Using the EHRead technology, different data operations/filters for inclusion and exclusion criteria will be applied to the target population to define the study population

Figure 1: **Study Design**



Index date
The index date is defined as the first inpatient record, or the second outpatient appointment record with a mention of MF diagnosis, within 6 months from the initial mention.

Study period
The study period will include a minimum **washout period** of 12 months (July 1, 2015, to June 30, 2016) and a **recruitment period** (incident cases, July 1, 2016, to end of study period).

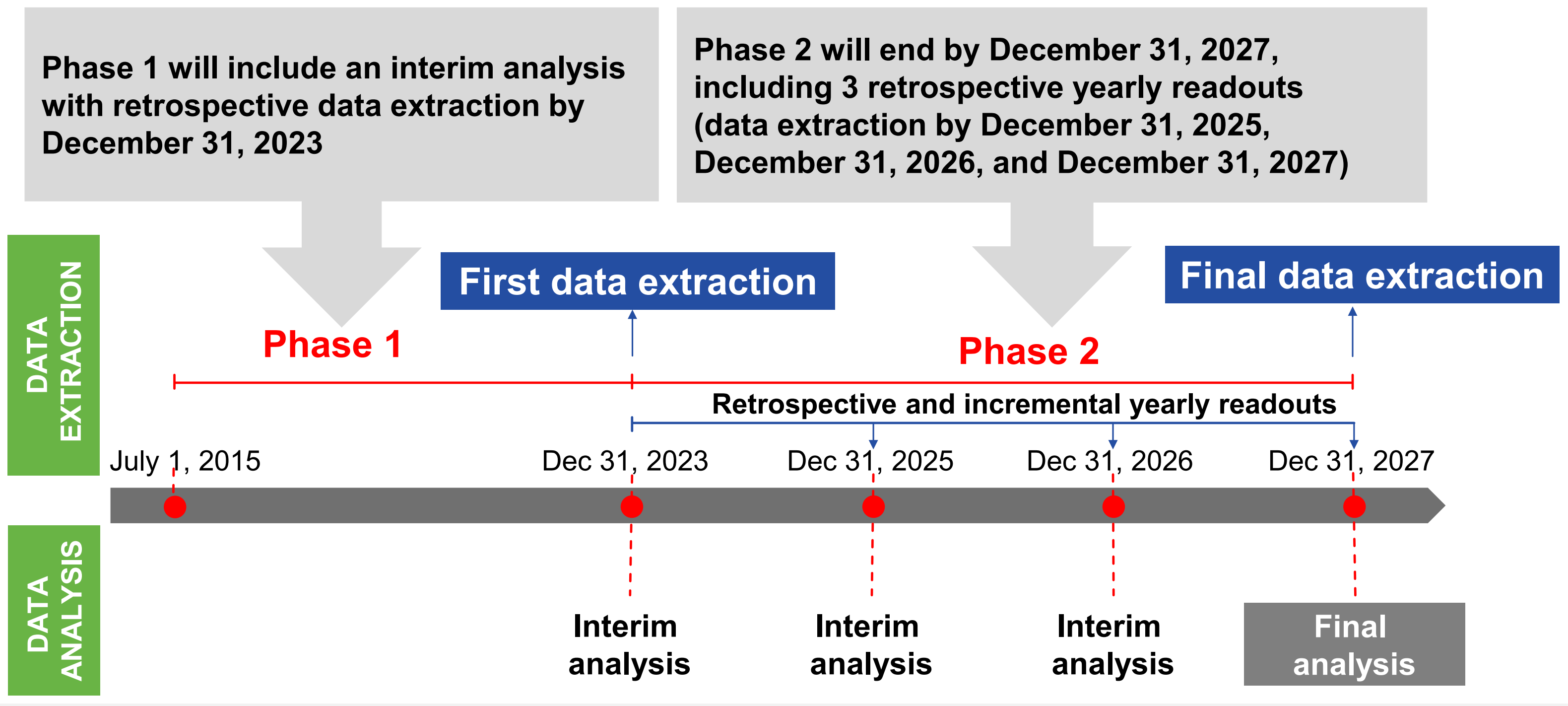
Inclusion criteria
- Primary/secondary MF diagnosis during recruitment period ≥6 months of preindex data

Exclusion criteria
- Age <18 years at diagnosis
- PV or ET diagnosis <6 months after the first MF diagnosis
- Allogeneic HSCT prior to the recruitment period[a]

Outcomes
- Demographics
- Comorbidities
- Clinical characteristics
- Treatment patterns
- Healthcare resource utilization
- Clinical outcomes
- Factors associated with disease progression and mortality

[a] HSCT criteria was applied to disambiguate potential misclassified MF diagnosis. Patients with an incident diagnosis of PV or ET after the first mention of MF or with the antecedent of allogeneic HSCT were excluded; patients with secondary MF and antecedent of PV or ET were included.
EHR, electronic health record; ET, essential thrombocythemia; HSCT, hematopoietic stem cell transplant; MF, myelofibrosis; ML, machine learning; NLP, natural language processing; PV, polycythemia vera.

Figure 2: **Data Extraction and Analysis Strategies**



Phase 1 will include an interim analysis with retrospective data extraction by December 31, 2023

Phase 2 will end by December 31, 2027, including 3 retrospective yearly readouts (data extraction by December 31, 2025, December 31, 2026, and December 31, 2027)

Results

- Currently, the study protocol has been approved by 7 participating hospitals, with approximately 229 patients to be included over an 11-year recruitment period
- Preliminary results are expected by June 2026
  - Current findings include the design of the terminology-based MF Disease Panel, internal evaluation of the cNLP for term extraction in Spanish, and early postprocessing and variable structuring design

1. MF Disease Panel

- The initial panel was developed in Spanish using 2 approaches:

1. **Expert review of scientific literature to identify key variables in MF (clinical characteristics, treatments, outcomes), with terms mapped to standard terminologies (SNOMED, ATC, LOINC)**
2. **Frequency analysis of clinical terms detected by our cNLP general processing pipeline, with frequent terms considered for inclusion**

- After the initial disease panel was created, it was refined through annotation projects. Medical annotators flagged synonyms missed by the NEL pipeline (false negatives) and removed incorrectly mapped terms (false positives)
  - Medical and terminology experts then reviewed the results and updated the panel by consensus
- The current MF Disease Panel contains 232 clinical terms
  - The average number of Spanish alternative expressions—such as alternative terms, abbreviations, and acronyms—per term was 5.8, with a range from 1 to 56
- The MF Disease Panel is ongoing for different languages (English, French, and German)

2. cNLP Performance Evaluation

Overall precision score
0.96

F1 score
0.94
Extraction of 41 terms in Spanish from the MF Disease Panel

Table 1: **cNLP Performance Metrics of 11 Representative Terms of the MF Disease Panel**

Term name	TP	FP	FN	Instances	Precision	Recall	F1
Myelofibrosis	602	28	31	633	0.96	0.95	0.95
Anemia	382	14	14	396	0.96	0.96	0.96
Splenomegaly	267	15	16	283	0.95	0.94	0.95
Fever	257	6	6	263	0.98	0.98	0.98
Ruxolitinib	358	22	26	384	0.94	0.93	0.94
Fedratinib	107	7	8	115	0.94	0.93	0.93
Hydroxycarbamide	289	14	14	303	0.95	0.95	0.95
Splenectomy	114	0	1	115	1.00	0.99	1.00
RBC transfusion	94	1	1	95	0.99	0.99	0.99
Disease progression	65	6	6	71	0.92	0.92	0.92
Acute myeloid leukemia	60	0	0	60	1.00	1.00	1.00

The number of instances was calculated as TP + FP + FN. Precision was calculated as TP/(TP + FP), indicating the accuracy of the information the system retrieves. Recall was calculated as TP/(TP + FN), indicating the amount of information the system retrieves. The F1 score was calculated as (2 × Precision × Recall)/(Precision + Recall) and is the overall performance indicator of information retrieval.
FN, false negative; FP, false positive; RBC, red blood cell; TP, true positive.

3. Postprocessing and Variable Structuring

- The postprocessing stage will transform the extracted clinical concepts, initially identified as raw terms, into research-ready variables
  - This includes variable-to-term mapping and temporal window alignment, ensuring that clinical events are correctly positioned within the appropriate time frames, and attribute modeling (eg, primary or secondary MF, score severity, Janus kinase inhibitor dosage); this work will ensure clinically interpretable, analyzable datasets that align with study objectives
  - The clinical data will be analyzed at different time points and windows, namely index date and follow-up period

Limitations

- There is inherent heterogeneity within EHR data, which are based on both structured and unstructured, free-text narratives written by healthcare professionals
  - Thus, the findings are limited by the accuracy of the documentation of patients' status, and the NLP's performance is limited by its ability to extract relevant information from text
- Only data explicitly present in EHRs (eg, free-text of structured fields) could be analyzed; missing, untested, or unrecorded variables (eg, unavailable laboratory results) could not be captured

Discussion and Conclusions

- **RWD studies are increasingly recognized as critical complements to randomized controlled trials in regulatory decision-making, with agencies such as the US Food and Drug Administration and European Medicines Agency encouraging their integration**
  - However, an estimated 80% of relevant information remains locked within unstructured free text,[3] limiting its usefulness for systematic analysis
- **cNLP provides a framework to extract, structure, and analyze this hidden information from unstructured text**
  - Transforming free text into standardized data enables comprehensive characterization of patients with MF, supporting generation of robust real-world evidence
- **Although cNLP evaluation is still ongoing, the models tested thus far demonstrated high performance, supporting their broader applicability in real-world oncology research and setting a foundation for future studies leveraging RWD to address complex clinical questions**

Abbreviations

ATC, Anatomical Therapeutic Chemical; cNLP, clinical Natural Language Processing; EHR, electronic health record; ET, essential thrombocythemia; FN, false negative; FP, false positive; HSCT, hematopoietic stem cell transplant; LOINC, Logical Observation Identifiers Names and Codes; MF, myelofibrosis; ML, machine learning; NEL, Named Entity Linking; NLP, natural language processing; PIO, patient, intervention, and outcome; PV, polycythemia veraRWD, real-world data; RBC, red blood cell; SLiCE, Smart Linguistic Corpus Extraction; SNOMED CT, Systematized Nomenclature of Medicine – Clinical Terms; TP, true positive.

References

1. Mughal TI, et al. *Int J Gen Med.* 2014;7:89-101.
2. Tefferi A. *Am J Hematol.* 2018;93(12):1551-1560.
3. Kong HJ. *Healthc Inform Res.* 2019;25(1):1-2.

Acknowledgments

This analysis was supported by GSK (study #221275).

Medical writing support was provided Keng Jin Lee, PhD, Nucleus Global, an Inizio company, and funded by GSK.

Disclosures

FSN, TdE, BV, PJ, JG, and SZ report employment at and stock or stock options with GSK; MP, ML, MG, LC-H, and BB are employees of Savana Research, which was hired by GSK to conduct this research.

ISPOR Europe 2025 | November 9-12, 2025 | Glasgow, Scotland, UK

Presenting author: Margarita Posso, mposso@savanamed.com