Using AI to Automatically Process Data from Unstructured Health Records of Patients with Lung Cancer

Mihaela Aldea[1], Pierre Rolland[2], Solenne Simon[1], Lodovica Zullo[1], Azeddine Djarallah[2], Aliette Poplu[2], Lisa Chuttoo[2], Muriel Wartelle[3], Benjamin Vignal[2], Jean-Charles Louis[2], François Lion[3], Arnaud Borie[2], David Planchard[1], Caroline Robert[1], Stefan Michiels[4], Fabrice Barlesi[1], Franck Le Ouay[2] , Benjamin Besse[1]

[1] Department of Medical Oncology, Gustave Roussy Cancer Center, Villejuif, France [2] Lifen, Paris, France [3] Informatic Team (DTNSI), Gustave Roussy, Villejuif, France [4] Biostatistics & Epidemiology, Gustave Roussy, Villejuif, France

lifen

GUSTAVE ROUSSY CANCER CAMPUS GRAND PARIS

P # 5084

BACKGROUND

- Structured databases created from electronic health records (EHR) are crucial for cancer research. Manual data entry into databases is both labor-intensive and error-prone.
- This study's objective was to create and validate an artificial intelligence (AI)-driven approach for automatically inputting lung cancer patient information from EHRs.

METHOD

POPULATION
Patients with thoracic cancer seen at Gustave Roussy between February 2021 and June 2023.

MANUAL DATA ENTRY (MDE)
Manual restrospective collection of data in a secured RedCap database.

AUTOMATED DATA ENTRY (ADE) – INPUT
- Unstructured patient medical letters between February 2021 - January 2024.
- A schematic description of each variable.

METHOD
- Generative AI to find, quote and process variables into a structured form.
- Large language model (LLM) actions with prompt engineering and tailored few-shots examples.
- Mortality data were auto-extracted from the French public registry, INSEE.

OUTPUT
Demographics, risk factors, molecular profile, cancer history, treatment data, survival data.

METRICS
Concordance between comparable dates from MDE and ADE, secondary manual review for mismatches (senior physician); correctness (accuracy after checking); time per patient.

RESULTS

Algorithm

INPUT
Patients | Cohort Variables

Lifen Documents
Doc 1 CLN | Doc 2 DR | Doc 3 AN

Variable Bank
Var 1 - Birthdate
Var 2 - Histological type
Var 3 - Diagnosis date

Document Processing
Doc 1 CLN | Doc 2 DR | Doc 3 AN
OCR | Remove Duplicates | PDF to Text

Variable Formatting
Var 1 - Birthdate
Var 1 - JSON Schema
Var 1 - Description, format, etc
Var 2 - Histological type
Var 2 - JSON Schema

LLM Engine
Output 1: JSON — Parse Json — Error management — Final Output: Exportable Data

Human verification and validation

OUTPUT
Structured Data

Concordance ADE vs MDE

PILOT STUDY

Advanced NSCLC treated with 1stl immunotherapy +/- chemo
N = 53 patients
n = 137 variables

Manual data entry N = 53 patients ? = Automated data entry N = 53 patients

7,261 Total data points
90.2% Concordance rate
6.8% Mismatch

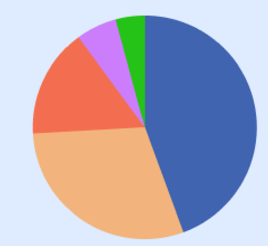
LUCC COHORT

Any thoracic cancer, any treatment received
N = 1057 patients
n = 206 variables

Manual data entry N = 137 patients ? = Automated data entry N = 137 patients
Automated data entry N = 920 patients

298,608 Total data points
92% Concordance rate
6.6% Mismatch

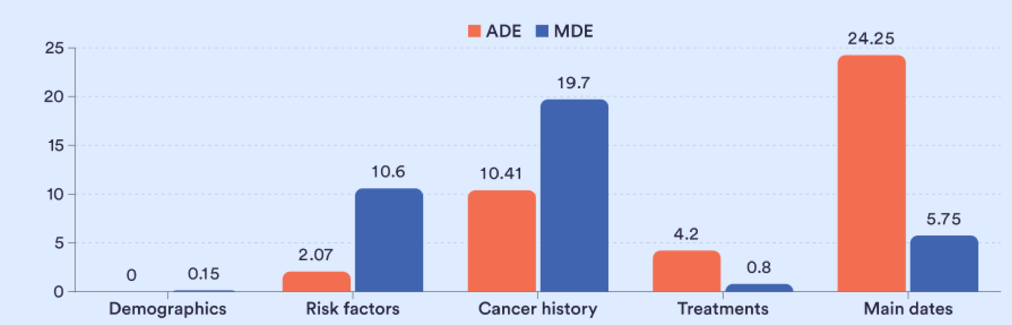
Mismatch



- 44.4% MDE errors
- 29.7% ADE errors due to limited access to data
- 15.8% ADE errors
- 5.9% MDE & ADE common errors
- 4.2% False errors

ADE errors were mostly from data gaps in medical notes. Detailed information was often accessible to MDE in imaging or pathology reports, yet inaccessible to ADE.

Missing Data



CORRECTNESS 95% - 100%
- Gender; Birthdate; Date of death
- Asbestos exposure
- COPD, myocardial infarction, autoimmune disease
- Histology
- Metastatic evolution anytime
- Systemic treatment class, drugs, sequence

90% - 94%
- Life status
- Smoking (yes/no)
- Dyslipidemia, diabetes
- Family history of lung cancer
- Metastatic from diagnosis
- Molecular alterations
- Sites of progression

80% - 89%
- Current/former smoking status
- Pack-years
- Thromboembolic event
- PDL1 score
- Metastatic sites at time of each systemic treatment
- Treatment discontinuation

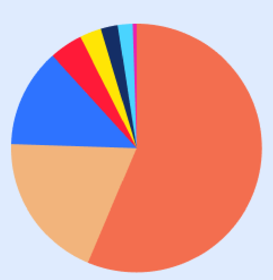
70 - 79%
- Cannabis consumption
- Joint-year
- Comorbidities: HTA
- Date diagnosis and first metastasis
- Stage cT, cN
- Date of start treatment

50 - 70%
- Date first metastasis
- Stage cM ;TMB value
- Start date of each treatment l
- Best objective response to each treatment l
- Event of progression to each treatment l

< 50%
- Date last follow-up for living patients
- ECOG performance status at each treatment start
- Date of last administration of treatment
- Date of progression for each treatment

ADE Cohort

Genomic alterations in 861 NSCLC - automated data entry



MOLECULAR
- 56.41% Other/none
- 19.2% KRAS
- 12.8% EGFR
- 4.3% ALK
- 2.8% RET
- 2.2% ROS1
- 1.9% BRAF
- 0.5% NTRK

Time per Patient

ADE **7 min** WITH QUALITY CHECK

MDE **>1 h**

CONCLUSION

- Generative AI can identify and structure unstructured data from EHRs, with >90% concordance between ADE and MDE.
- High performance of ADE is seen with demographics, risk factors, comorbidities, histology, molecular profile and treatment types, while lower performance with dates (e.g. last follow-up, progression or last scan without progression).
- ADE cases of low correctness are often due to a lack of information in medical notes.
- ADE has the potential to enhance the efficiency, accuracy, and scalability of EHR-to-database conversions.