Article

A Natural Language Processing Algorithm to Improve Completeness of ECOG Performance Status in Real-World Data

Aaron B. Cohen [1,2,*], Andrej Rosic [1], Katherine Harrison [1], Madeline Richey [1], Sheila Nemeth [1], Geetu Ambwani [1], Rebecca Miksad [1], Benjamin Haaland [3] and Chengsheng Jiang [1]

[1] Flatiron Health Inc., 233 Spring St., New York, NY 10013, USA
[2] Department of Medicine, NYU Grossman School of Medicine, New York, NY 10016, USA
[3] Huntsman Cancer Institute, University of Utah, Salt Lake City, UT 84112, USA
* Correspondence: acohen@flatiron.com

Supplemental Materials

Table S1. Detailed cohort eligibility criteria

Table S2. Impact of the application of the NLP algorithm on the ECOG PS completeness in EHR-derived databases for 21 diseases.

Table S3. rwOS (months) in patients with aNSCLC stratified according to their ECOG PS score for the subcohort with ECOG PS scores available as structured data, and for the subcohort with ECOG PS scores extracted via algorithm.

TableS4. HR for patients with structured ECOG (reference group) in patients with aNSCLC

Table S5. Median real world OS (months) for patients present in the testing set across all eligible diseases.

Figure S1. rwOS in patients present in the study databases across all eligible diseases, stratified according to their ECOG PS score.

Table S1. Detailed cohort eligibility criteria. All patients were required to have at least two documented clinical visits on separate days on or after January 1, 2011, except for patients with mCRC, prostate cancer (metastatic), or SCLC, who had to have at least two documented clinical visits on separate days on or after January 1, 2013, and patients with pancreatic cancer (metastatic), who had to have at least two documented clinical visits on separate days on or after January 1, 2014. Additionally, the following inclusion/exclusion criteria for each cancer type are as follows:

Cancer Type	Inclusion criteria			Exclusion criteria
	ICD codes	Pathology and staging	Additional	
eBC	Diagnosed with breast cancer (ICD-9 174.x or 175.x or ICD-10 C50.x)	Pathology consistent with breast cancer; Has evidence of stage I - III breast cancer with a diagnosis date on or after January 1, 2011.	NA	NA
mBC	Diagnosed with breast cancer (ICD-9 174.x or 175.x or ICD-10 C50x)	Pathology consistent with breast cancer; Has evidence of stage IV or recurrent metastatic breast cancer with a metastatic diagnosis date on or after January 1, 2011.	NA	NA
CLL	Diagnosed with CLL (ICD-9: 204.1x or ICD-10: C91.1x, C83.0x)	Physician documentation of CLL; Has evidence in unstructured documents of having been treated specifically for CLL.	At least one order for an antineoplastic occurring on or after January 1, 2011	NA
mCRC	Diagnosed with CRC cancer (ICD-9 153.x	Pathology consistent with CRC; Has evidence of Stage IV or recurrent metastatic	NA	NA

	or 154.x or ICD-10 C18x, or C19x, or C20x, or C21x)	CRC diagnosed on or after January 1, 2013.		
DLBCL	Diagnosed with Non-Hodgkin's Lymphoma (ICD 9: 200x, 202x; ICD 10: C82x, C83x, C84x, C85x, C86x, C88x, C96x)	Has evidence of DLBCL with an initial diagnosis date on or after January 1, 2011	NA	NA
aGE	Diagnosed with Advanced Gastric/Esophageal: Gastric: ICD-9 code 151-151.9 (151.x) or ICD-10 code C16-C16.9 (C16.x); GEJ/Esophageal: ICD-9 code 150-150.9 (150.x) or ICD-10 code C15-C15.9 (C15.x)	Pathology consistent with Gastric/Esophageal cancer; For gastric cancer: Patients with stage IV disease at diagnosis or with one of the following: 1) distant recurrence, 2) a 2nd locoregional recurrence 3) a 1st locoregional recurrence that was not completely resected or 4) no surgical resection of the primary tumor; Diagnosis of advanced disease as described above must be on or after January 1, 2011; For esophageal/gastroesophageal junction cancer: Patients with stage IV disease at diagnosis or with one of the following: 1) distant recurrence, 2) any locoregional recurrence or 3) no surgical resection of the primary tumor Diagnosis of advanced disease as described above must be on or after January 1, 2011	NA	NA
HCC	Diagnosed with HCC (ICD-9 155.x or ICD-10 C22.x)	Pathology consistent with HCC (mixed types excluded); Diagnosed with HCC on or after January 1, 2011.	NA	NA
aHNC	Diagnosed with head & neck cancer (ICD 9: 140x, 141x,	Pathology consistent with squamous cell carcinoma of the Head & Neck; Diagnosed with advanced Head & Neck SCC on or	Primary site of disease in the Oral Cavity, Oropharynx,	Histology other than squamous cell carcinoma;

	143x, 144x, 145x, 146x, 147x, 148x, 149x, 161x; ICD 10: C00x, C01x, C02x, C03x, C04x, C05x, C06x, C09x, C10x, C11x, C12x, C13x, C14x, C32x	after 1/1/2011.	Hypopharynx, Pharynx NOS, Larynx, or Unknown primary Head & Neck cancer	Primary site of Head & Neck cancer other than the primary sites listed.
aMel	Diagnosed with melanoma (ICD-9 172.x or ICD-10 C43x or D03x)	Patients with pathologic stages III or IV at initial diagnosis on or after January 1, 2011, or patients with earlier stage disease who develop a first locoregional or distant recurrence on or after January 1, 2011	NA	Non-skin melanoma (ocular, subungual, mucosal, palmar, plantar); Stage I-II melanoma without a locoregional or distant recurrence; Diagnosis before January 1, 2011 and recurrence date prior to January 1, 2011
MM	Diagnosed with multiple myeloma (ICD-9 203.0x or ICD-10 C90.0x, C90)	Pathology consistent with Multiple Myeloma; Has evidence of multiple myeloma with a diagnosis date on or after January 1, 2011	NA	NA
MPM	Diagnosed with MPM (ICD-9 163* or ICD-10 C45.0)	Pathology consistent with MPM; Diagnosed with MPM on or after January 1, 2011	NA	NA
aNSCLC	Diagnosed with lung cancer (ICD-9 162.x	Pathology consistent with NSCLC; Diagnosed with Stage IIIB, IIIC, IVA or IVB NSCLC	NA	NA

	or ICD-10 C34x or C39.9)	on or after 1/1/2011, or diagnosed with early-stage NSCLC and subsequently develops recurrent or progressive disease on or after 1/1/2011.		
Ovarian	ICD code for ovarian, fallopian tube, and/or peritoneal cancer (ICD 9: 183x, 158x; ICD 10: C56x, C57.0x, C48x)	Diagnosed with invasive ovarian, fallopian tube, and/or primary peritoneal cancer on or after 1/1/2011; Histology of one of the following: Serous, Mucinous, Clear Cell, Transitional Cell, Endometrioid, Epithelial NOS, Borderline, Unknown/not documented.	NA	Gender = Male; Histology other than the histologies listed (e.g. Squamous cell, Germ cell/ theca cell/granulosa cell, Sex cord stromal, Dysgerminoma, Teratoma (mature or immature), Yolk sac tumor, MMMT (malignant mixed mullerian tumor), Benign or malignant Brenner tumor, Sarcoma (usually leiomyosarcoma), Pseudomyxoma peritoneii, Mixed tumors with one or more non-epithelial components)
Pancreatic (met)	ICD code for pancreatic cancer (ICD-9 157.x; ICD-10 C25.x)	Pathology consistent with adenocarcinoma of the pancreas; Diagnosed with Stage IV disease on or after 1/1/2014 or diagnosed with earlier-stage pancreatic cancer and subsequently developed recurrent or progressive disease on or after 1/1/2014.	NA	Histology other than adenocarcinoma of the pancreas
Prostate (met)	Diagnosed with Metastatic Prostate Cancer: ICD 9 code: 185x or	Diagnosis of metastatic disease as described above must be on or after January 1, 2013; Histology is confirmed as adenocarcinoma or NOS (Not Otherwise Specified)	NA	Gender = Female or Unknown

	ICD 10 code: C61	Has a known CRPC or HSPC status.		
mRCC	Diagnosed with RCC (ICD-9 189.x or ICD-10 C64x or C65x)	Pathology consistent with RCC; Has evidence of stage IV or recurrent metastatic RCC with a Stage IV initial diagnosis date or metastatic diagnosis date on or after January 1, 2011.	NA	NA
SCLC	Diagnosed with lung cancer (ICD-9 162.x or ICD-10 C34x, or C39.9)	Pathology consistent with SCLC; Diagnosed with SCLC on or after 1/1/2013.	NA	Diagnosed with NSCLC on or before the time a patient was first abstracted for the SCLC cohort. Note that if a patient subsequently develops NSCLC after having entered the SCLC cohort, that patient will not be removed from the SCLC cohort
Urothelial (adv)	Diagnosed with urothelial cancer (ICD-9 188x, 189.1, 189.2, 189.3, or ICD-10 C65x, C66x, C67x, C68.0)	Pathology consistent with transitional cell (urothelial) carcinoma; Diagnosed with Stage IV urothelial carcinoma or node positive urothelial carcinoma on or after 1/1/2011, or diagnosed with early-stage urothelial carcinoma and subsequently develops advanced disease on or after January 1, 2011.	NA	Histology other than transitional cell (urothelial) carcinoma; Primary site of disease other than bladder, renal pelvis, ureter, or urethra.

e (m) BC=early (metastatic) breast cancer; CLL=chronic lymphocytic leukemia; mCRC=metastatic colorectal cancer; DLBCL=diffuse large B-cell lymphoma; aGE=advanced gastroesophageal; HCC=Hepatocellular carcinoma; aHNC=advanced head and neck cancer; ICD=International Classification of Diseases; MM=multiple myeloma; MPM=malignant pleural mesothelioma; NA=not applicable/available; aNSCLC=advanced non-small cell lung cancer; mRCC=metastatic renal cell carcinoma; SCLC=small-cell lung cancer

Table S2. Availability of ECOG PS Scores (percent of patients) via different sources across tumor types and lines of therapy in the study databases

Disease	**ECOG PS source**	**1L**	**2L**	**3L**	**4L**	**≥5L**
AML	Structured data	37.5	46.1	43.4	39.8	35.4
	Extracted by NLP algorithm	12.3	18.9	20.5	20.7	20.2
	After algorithm applied (structured +extracted)	49.8	65.0	63.9	60.5	55.6
Urothelial	Structured data	64.3	69.0	70.3	69.7	71.0
	Extracted by NLP algorithm	12.5	11.1	9.5	11.6	11.2
	After algorithm applied (structured +extracted)	76.8	80.1	79.8	78.7	82.2
metBrCa	Structured data	50.4	60.1	63.7	67.1	69.5
	Extracted by NLP algorithm	12.0	11.9	11.5	11.6	11.0
	After algorithm applied (structured +extracted)	65.5	72.0	75.2	78.7	80.5
CLL	Structured data	55.4	60.9	62.5	63.9	62.6
	Extracted by NLP algorithm	10.1	9.8	8.8	10.0	7.6
	After algorithm applied (structured +extracted)	65.5	70.7	71.3	73.9	70.2
CRC	Structured data	63.0	70.0	72.4	71.8	72.6
	Extracted by NLP algorithm	11.8	10.7	10.9	12.2	12.4
	After algorithm applied (structured +extracted)	74.8	80.7	83.3	84.0	85.0
DLBCL	Structured data	44.0	50.2	51.8	50.9	51.8
	Extracted by NLP algorithm	11.9	14.3	15.1	17.3	16.8
	After algorithm applied (structured +extracted)	55.9	64.5	66.9	68.2	68.6
eBrCa	Structured data	58.9	67.2	69.9	66.4	71.0
	Extracted by NLP algorithm	10.3	10.1	9.3	78.8	10.8
	After algorithm applied (structured +extracted)	69.2	77.3	79.2	12.4	81.8
Endometr.	Structured data	56.9	60.2	60.9	60.9	63.5
	Extracted by NLP algorithm	15.7	15.8	17.5	18.8	15.2
	After algorithm applied (structured +extracted)	72.6	76.0	78.4	79.7	78.7
FL	Structured data	48.0	56.2	58.3	59.4	54.6
	Extracted by NLP algorithm	10.9	11.5	12.1	11.4	13.7
	After algorithm applied (structured +extracted)	58.9	67.7	70.4	70.8	68.3
Gastric	Structured data	62.6	67.4	70.2	69.4	69.6
	Extracted by NLP algorithm	13.2	12.9	13.1	13.4	11.0
	After algorithm applied (structured +extracted)	75.8	80.3	83.3	82.8	80.6
HCC	Structured data	55.6	70.0	67.4	73.3	84.2
	Extracted by NLP algorithm	16.1	12.1	12.3	13.8	5.3
	After algorithm applied (structured +extracted)	71.7	82.1	79.7	87.1	89.5
Head Neck	Structured data	66.4	71.7	72.1	72.4	69.7
	Extracted by NLP algorithm	12.3	11.3	11.7	12.4	11.8
	After algorithm applied (structured +extracted)	78.7	83.0	83.8	84.8	81.5
MCL	Structured data	44.6	52.2	55.4	61.3	62.6
	Extracted by NLP algorithm	13.7	13.3	13.3	13.8	13.9
	After algorithm applied (structured +extracted)	58.3	65.5	68.7	75.1	76.5

aMelanoma	Structured data	58.0	55.0	55.8	51.0	46.2
	Extracted by NLP algorithm	19.6	24.7	27.8	32.7	38.8
	After algorithm applied (structured +extracted)	77.6	79.7	83.6	83.7	85.0
MM	Structured data	45.8	53.0	58.4	61.3	64.6
	Extracted by NLP algorithm	14.9	14.7	14.3	13.6	15.1
	After algorithm applied (structured +extracted)	60.7	67.7	72.7	74.9	79.7
aNSCLC	Structured data	62.9	65.4	66.7	67.3	64.6
	Extracted by NLP algorithm	13.6	13.0	13.5	14.0	15.1
	After algorithm applied (structured +extracted)	76.5	78.4	80.2	81.3	79.7
Ovarian	Structured data	58.4	63.9	66.4	67.3	69.7
	Extracted by NLP algorithm	11.5	11.7	12.7	12.5	12.4
	After algorithm applied (structured +extracted)	69.9	75.6	79.1	79.8	82.1
Pancreatic	Structured data	68.6	63.9	66.4	67.3	69.7
	Extracted by NLP algorithm	11.5	11.7	12.7	12.5	12.4
	After algorithm applied (structured +extracted)	69.9	75.6	79.1	79.8	82.1
PrCa	Structured data	55.0	60.9	66.5	67.8	68.9
	Extracted by NLP algorithm	11.8	12.0	12.0	13.3	14.0
	After algorithm applied (structured +extracted)	66.8	72.9	78.5	81.1	82.9
RCC	Structured data	54.2	59.8	63.1	63.7	66.8
	Extracted by NLP algorithm	14.4	14.8	16.0	15.8	16.4
	After algorithm applied (structured +extracted)	68.6	74.6	79.1	79.5	83.2
SCLC	Structured data	64.6	72.8	71.9	72.2	62.5
	Extracted by NLP algorithm	11.0	9.5	11.9	11.2	11.5
	After algorithm applied (structured +extracted)	75.6	82.3	83.8	83.4	74.0

Table S3: Real world median overall survival rwOS (months) in patients with aNSCLC stratified according to their ECOG PS score for the subcohort with ECOG PS scores available as structured data, and for the subcohort with ECOG PS scores extracted via algorithm.

Median Overall Survival (Months)		
ECOG PS	**Extracted (CI)**	**Structured(CI)**
0	18.8(17.5, 20.6)	18.4(17.7, 19)
1	12.7(12, 13.7)	11.9(11.6, 12.3)
2	6.6(6.2, 7.4)	6.7(6.4, 7)
3	4.8(4.2, 6.2)	4.1(3.6, 4.5)

Table S4: HR for patients with structured ECOG PS (Ref: NLP-Extracted) in patients with aNSCLC

Hazard Ratios for Patients with Structured ECOG PS (Ref: NLP-Extracted)

ECOG PS	HR	lowCI	upperCI
0	1.14	0.99	1.3
1	1.03	0.96	1.1
2	1.06	1.02	1.11
3	1.05	0.99	1.12

Table S5: Median real world OS (months) for patients present in testing sets across all eligible diseases

Median Overall Survival (Months)		
ECOG PS	Label-Data (CI)	Extracted (CI)
0	26.7(23.7, 31)	26.4(23.7, 30.8)
1	13.4(12.3, 14.9)	13.3(12.1, 14.6)
2	7(5.9, 9.1)	6.7(5.8, 8.4)
3	5.8(3.3, 8.8)	4.3(2.8, 7.6)

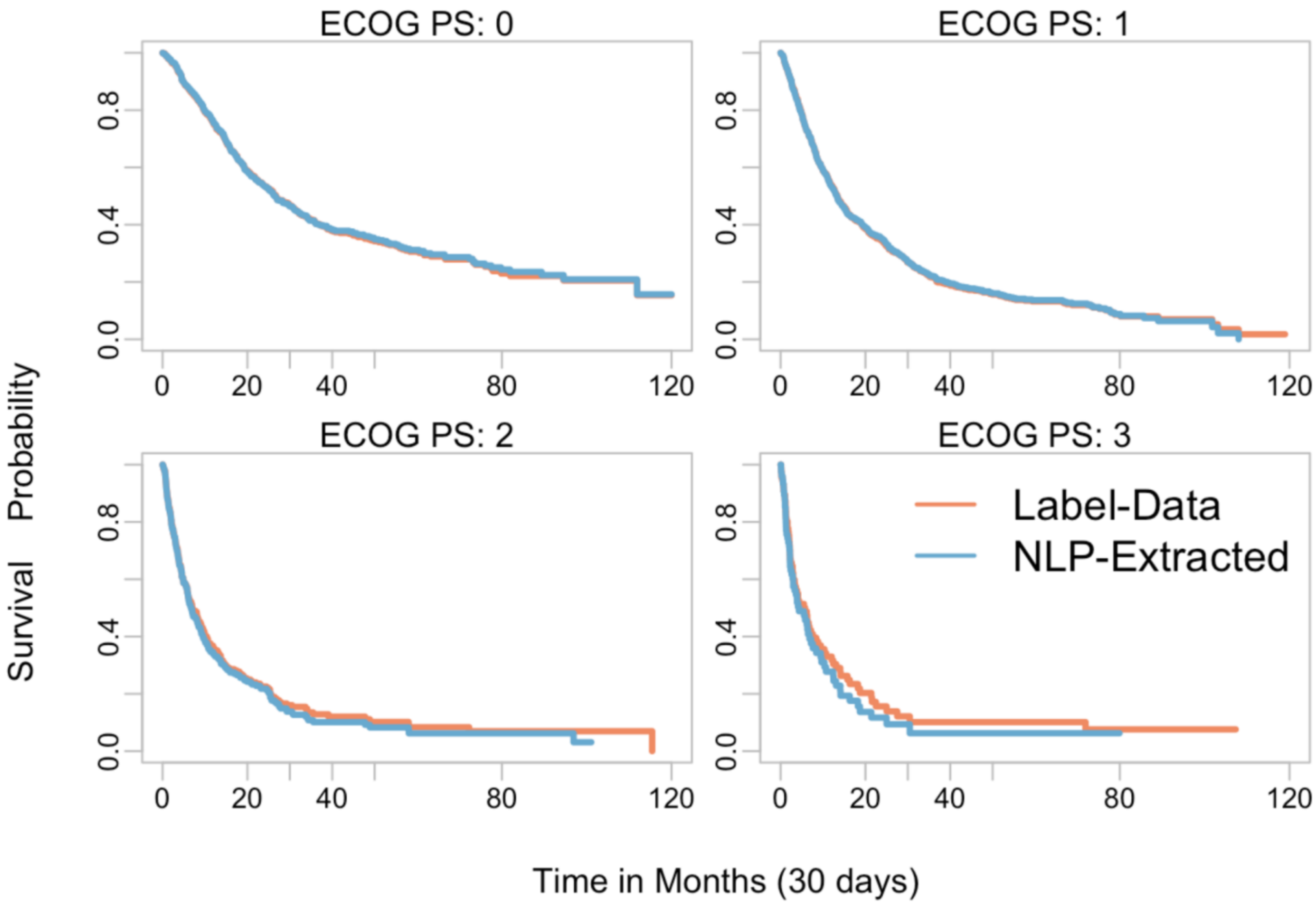

Figure S1. rwOS in patients present in the testing set across all eligible diseases, stratified according to their ECOG PS score, for the subcohort with ECOG PS scores in duplicate, as abstracted information, and as extracted via algorithm.