Supplementary Online Content

Lott JP, MD, Boudreau DM, Barnhill RL, et al. Population-based analysis of histologically confirmed melanocytic proliferations using natural language processing. *JAMA Dermatol*. Published online November 1, 2017. doi:10.1001/jamadermatol.2017.4060

eAppendix 1. Skin Biopsy Identification and Melanoma Identification Using Corresponding HCPCS/CPT 4, ICD-9, and ICD-O-3 Codes

eAppendix 2. Additional Details on NLP Methods and Health Plan SEER Data

eAppendix 3. Diagnosis at patient level at the index biopsy and when the patient record was followed forward to 90 days and 365 days to evaluate follow-up biopsy results. Results are shown overall and stratified by sex.

This supplementary material has been provided by the authors to give readers additional information about their work.

Population-Based Analysis of Histologically Confirmed Melanocytic Proliferations Using Natural Language Processing

Supplementary Electronic Appendix

Appendix Contents:

APPENDIX 1. Skin Biopsy Identification and Melanoma Identification Using

Corresponding HCPCS/CPT 4, ICD-9, and ICD-O-3 Codes

1A. HCPCS / CPT4 codes identifying skin biopsies and associated definitions

11100: BIOPSY SKIN LESION
11101: BIOPSY SKIN ADD-ON
11755: BIOPSY NAIL UNIT
40490: BIOPSY OF LIP
41100: BIOPSY OF TONGUE
41105: BIOPSY OF TONGUE
41108: BIOPSY OF FLOOR OF MOUTH
54100: BIOPSY OF PENIS
54105: BIOPSY OF PENIS
56605: BIOPSY OF VULVA/PERINEUM
56606: BIOPSY OF VULVA/PERINEUM
67810: BIOPSY EYELID & LID MARGIN
69100: BIOPSY OF EXTERNAL EAR

1B. ICD 9 procedure code identifying skin biopsies and associated definition
86.1: Skin & subcutaneous diagnostic procedure

1C. In the automated data, melanoma cases were identified using histology codes and behavior codes 2 = *in situ* and 3 = invasive melanoma. The ICD-O-3 histology codes and definitions are shown below

ICD-O-3 Code	Histology Name
8720/2	Melanoma in situ
8720/3	Malignant melanoma, NOS
8721/3	Nodular melanoma
8723/3	Malignant melanoma, regressing
8730/3	Amelanotic melanoma
8740/3	Malignant melanoma in junctional naevus
8741/2	Precancerous melanosis, NOS
8741/3	Malignant melanoma in precancerous melanosis
8742/2	Hutchinson's melanotic freckle, NOS
8742/3	Malignant melanoma in Hutchinson's melanotic freckle
8743/3	Superficial spreading melanoma
8744/3	Acral lentiginous melanoma, malignant
8745/3	Desmoplastic melanoma, malignant
8761/3	Malignant melanoma in giant pigmented naevus
8770/3	Mixed epithelioid and spindle cell melanoma
8771/3	Epithelioid cell melanoma
8772/3	Spindle cell melanoma, NOS

8773/3	Spindle cell melanoma, type A
8774/3	Spindle cell melanoma, type B

Appendix 2. Additional Details on NLP Methods and Health Plan SEER Data

Appendix 2A. NLP Methods

Original pathology reports (N=289) were independently reviewed and classified into the MPATH-Dx system by two experienced dermatologists (JL, EK) and any disagreements reviewed in consensus using a modified Delphi method.[34]

A string search method was initially used to try and extract the information from each pathology report. We then used the phrases as a jumping off point and created a simple context free grammar that generated 6,455 different phrases, all linked back to their associated MPATH-Dx class by the name of the grammar rule that licensed it. For example:

S -> A N

N -> ‘nevus’

A -> 1N | 2N

1N -> ‘blue’ | ‘junctional’

2N -> ‘spitz’

In this simplified version of the grammar, a “sentence” must meet the rule “S”, meaning it must be made up of an “A” and an “N”. An “N” can only be the word “nevus”. An “A” can be a “1N” or a “2N”. A “1N” can be “blue” or “junctional”. A “2N” can be “spitz”. Therefore, this simple grammar would generate the phrases “blue nevus”, “junctional nevus” and “spitz nevus”, and all can be linked back to their MPATH-Dx class by the rule that licensed them. “1N” words are of MPATH-Dx class 1, and “2N” words are of MPATH-Dx class 2. The grammar that we used in the NLP pipeline is much larger than this and allows for more variation in word order, but the basic principle is the same. The systems first step is to read in the grammar and create a dictionary of phrases mapped to their MPATH-Dx class.

As each phrase is created, it is turned into a regular expression that allows for flexible spelling and spacing between words. Spacing tends to vary within text of the pathology reports, and, if the system were to assume single spacing it would miss a fair amount of instances. The regular expression allows for 0 or more spaces between words, to account for typographical errors. Regular expressions also allow for faster searching and matching than strings.

The second step was to read in the rules for the modified version of the NegEx algorithm.[29] There are two types of rules here: “linking” rules and “negation” rules. The

linking rules describe conjunctions, such as “and”, “or”, and commas. These are used by the algorithm to ensure that phrases that are linked are both negated, such as in “no melanocytes or nevus detected”. Both “melanocytes” and “nevus” are negated here, and that should be reflected. The negation rules are of the type “PREN” or “POST”, signifying pre or post phrase negation. Pre-negations appear before the phrase in question and post-negations appear after the phrase in question. These are generally simple phrases, such as “no evidence of”, “markedly declined” or simply the word “no”. The algorithm allows for some scope, and, if a negation phrase occurs within two words or punctuation marks of the phrase, then the phrase is considered negated.

The pathology reports are read into the system once the phrases are generated and the negation algorithm is trained. Because each pathology report can have multiple diagnoses from multiple biopsy sites, the system first separates these. Biopsy sites are extracted from each report as the report is created. So, if the text says “A) blue nevus found on the elbow. B) no melanocytic lesions on the thigh”, then those two sites are separated out and given different MPATH-Dx classes (A would be a class 1 and B would be a class 0). Each report has a unique label, a date, a patient ID, a list of the biopsy sites and their MPATH-Dx classes, a “most severe” field that links to the most severe diagnosis the report contained, and the raw text of the report.

The biopsies are classified as -1, 0, 1, 2, 3, 4,5, or 6 to represent the non-melanocytic lesions and the MPATH-Dx classification tool as shown in **Table 1**. Basically, NLP Level -1 = cases where the NLP was not able to classify, NLP Level 0 = skin biopsies that had tissue with no melanocytic lesions and NLP Levels 1 and above were skin biopsies with varying types of melanocytic lesions.

We did not train the NLP system to extract the severity of invasive melanomas, which is the purpose of MPATH-Dx classes 4 and 5. Instead we are able to use SEER data to further classify the NLP Level 6 invasive melanoma diagnoses that the NLP system was not able to classify into the expected proportions that would be found within MPATH-Dx Class 4 and 5. Because NLP levels 4 and 5 are actually groupings of several different severity levels of invasive melanoma and because the NLP system judges severity through integer comparison, a more discrete numbering system was used and is presented in the table below. Thus NLP Levels 4 to 9 represent all of the invasive melanoma cases.

eTable. Explanation of NLP levels as they relate to clinical diagnoses and the MPATH-Dx classification system.

NLP Level	Clinical Comments and Example Diagnoses	MPATH-Dx Class
-1	Unclassifiable by NLP system	N/a
0	No melanocytic lesion noted (e.g. inflammatory disease, squamous cell carcinoma, etc.)	0

1	Mild atypical nevus and other similar diagnoses	I
2	Moderately atypical nevus and other similar diagnoses	II
3	Melanoma *in situ*, severely atypical nevi	III
4	pT1a invasive melanoma	IV
5	pT1b invasive melanoma	V
6	Invasive melanoma, not able to classify	V
7	T2 invasive melanoma	V
8	T3 invasive melanoma	V
9	T4 invasive melanoma	V

The NLP system reviews data by patient, but, because a patient may have more than one entry, a more nuanced approach is required. It was found in an earlier study[21] that pathology reports and diagnosis can be reliably linked to their original procedure report using the patient ID and a +/- 14 day date window. If the biopsy in question is classified as a 6, the system searches for the patient ID first, then searches for a diagnosis that was created within 14 days of the pathology report being entered into the health system EMR database. If no further diagnosis is found meeting those criteria, then the classification is left as a 6.

If the patient ID has not been seen before, then a new patient is created. Each patient object in memory contains the date of first index biopsy, the total number of biopsies the patient has received. For the analysis at the patient level showing diagnostic drift over the subsequent 365 days, the most severe diagnosis the patient received during a one-year period after the index biopsy was noted, and an ordered timeline of all the skin biopsy pathology reports associated with the patient. This timeline is updated when a new pathology report is added. The patient objects are also able to create a snapshot of the patient's biopsy history in a given timeframe, assuming that the index date is the date of the first biopsy. This query process can be used to see the data in different ways (e.g., describing the highest level diagnosis when considering all biopsies within 30 days of the index biopsy, within 90 days of the index biopsy, etc).

As we are able to estimate the percentage of invasive melanoma within MPATH-Dx class IV vs. V from health plan SEER data, we did not pursue the process of using the NLP system to differentiate the invasive melanoma MPATH-Dx class IV vs. class V cases.

Appendix 2B. Summary of SEER[1] Data and Method Used to Estimate the Percentage of cases in MPATH-Dx Class IV vs. V.

The NLP search of pathology reports was able to distinguish AJCC primary tumor stage t1a vs. >=t1b invasive melanomas for some, but not all cases. In order to estimate the percentage of invasive melanoma that were MPATH-DX Class IV (pt1a) vs MPATH-Dx Class V (>=pt1b) we used local health plan SEER data.[1] Projections below used SEER data for both men and women > 18 years of age at the time of the diagnosis

The percent of all invasive melanomas computed from the health plan SEER were 52.8% in pathologic stage p1a and 47.2% in pathologic stage t1b or higher. Primary tumors that could not be assessed (e.g., T1 NOS or missing T N=98) were not considered in these estimates.

Invasive melanoma diagnosis frequency counts among male and female patients from the tumor registry (2007-2012), and estimates of the percent that fall into the MPATH-Dx Class IV vs V. These total numbers are slightly different from the total invasive melanoma diagnoses shown in Table 1 of the paper as there was no requirement of one year of continuous enrollment for eligibility as in our study and this includes all cases of invasive melanoma from the full 2007 (when some of the pathology reports were not included in the NLP)

eTable. Classification of Invasive Melanoma Cases Using Health Plan SEER Tumor Registry Data

MPATH-Dx Class	AJCC Primary Tumor Stage[2]	Count	Column %
Class IV	Invasive Melanoma (T1a)	437	52.8%
Class V	Invasive Melanoma (T1b -T4b)	390	47.2%
	Total	827	100%

References:

1. Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) SEER*Stat Database: Incidence - SEER 18 Regs Research Data, Nov 2015 Sub (1973-2013) <Katrina/Rita Population Adjustment> - Linked To County Attributes - Total U.S., 1969-2014 Counties, National Cancer Institute, DCCPS, Surveillance Research Program, Surveillance Systems Branch, released April 2016, based on the November 2015 submission.
2. Edge SB, Byrd DR, Compton CC, Fritz AG, Greene FL, Trotti A, (editors). ***AJCC cancer staging manual, 7th edition.*** France: Springer; 2010 [cited 2017 Jan 25] Available from: http://www.springer.com/medicine/surgery/book/978-0-387-88440-0.

Appendix 3. Diagnosis at patient level at the index biopsy and when the patient record was followed forward to 90 days and 365 days to evaluate follow-up and biopsy results[a]. Results are shown overall and stratified by sex.

Overall

MPATH-Dx Class	18-44 years			45-64 years			≥ 65 years		
	Index n=4723 (%)	90 days n=4723 (%)	365 days n=4724 (%)	Index n=4550 (%)	90 days n=4555 (%)	365 days n=4569 (%)	Index n=3443 (%)	90 days n=3451 (%)	365 days n=3467 (%)
I	4192 (89%)	4176 (88%)	4160 (88%)	3904 (86%)	3897 (86%)	3901 (85%)	2661 (77%)	2644 (77%)	2648 (76%)
II	397 (8%)	402 (9%)	416 (9%)	373 (8%)	377 (8%)	385 (8%)	299 (9%)	284 (8%)	282 (8%)
III	85 (2%)	94 (2%)	96 (2%)	134 (3%)	137 (3%)	138 (3%)	232 (7%)	248 (7%)	256 (7%)
IV and V	49 (1%)	51 (1%)	52 (1%)	139 (3%)	144 (3%)	145 (3%)	251 (7%)	275 (8%)	281 (8%)

[a]Some columns do not add up to 100% due to rounding.

Men

MPATH-Dx Class	18-44 years			45-64 years			≥ 65 years		
	Index N (%)	90 days N (%)	365 days N (%)	Index N (%)	90 days N (%)	365 days N (%)	Index N (%)	90 days N (%)	365 days N (%)
I	2905 (89%)	2895 (89%)	2881 (88%)	2517 (88%)	2513 (88%)	2517 (88%)	1430 (82%)	1427 (81%)	1427 (81%)
II	263 (8%)	265 (8%)	277 (8%)	192 (7%)	194 (7%)	199 (7%)	115 (7%)	110 (6%)	108 (6%)
III	66 (2%)	73 (2%)	75 (3%)	68 (2%)	69 (2%)	69 (2%)	93 (5%)	97 (6%)	103 (6%)
IV and V	35 (1%)	36 (1%)	37 (1%)	72 (3%)	75 (3%)	75 (3%)	110 (6%)	119 (7%)	124 (7%)
Total	3269 (100%)	3269 (100%)	3270 (100%)	2849 (100%)	2851 (100%)	2860 (100%)	1748 (100%)	1753 (100%)	1762 (100%)

[a]Some columns do not add up to 100% due to rounding.

Women

MPATH-Dx Class	18-44 years	45-64 years	≥ 65 years

	Index N (%)	90 days N (%)	365 days N (%)	Index N (%)	90 days N (%)	365 days N (%)	Index N (%)	90 days N (%)	365 days N (%)
I	1287 (89%)	1281 (88%)	1279 (88%)	1387 (82%)	1384 (81%)	1384 (81%)	1231 (73%)	1217 (72%)	1221 (72%)
II	134 (9%)	137 (9%)	139 (10%)	181 (11%)	183 (11%)	186 (11%)	184 (11%)	174 (10%)	174 (10%)
III	19 (1%)	21 (1%)	21 (1%)	66 (4%)	68 (4%)	69 (4%)	139 (8%)	151 (9%)	153 (9%)
IV and V	14 (1%)	15 (1%)	15 (1%)	67 (4%)	69 (4%)	70 (4%)	141 (8%)	156 (9%)	157 (9%)
Total	1454 (100%)	1454 (100%)	1454 (100%)	1701 (100%)	1704 (100%)	1709 (100%)	1695 (100%)	1698 (100%)	1705 (100%)

[a]Some columns do not add up to 100% due to rounding.