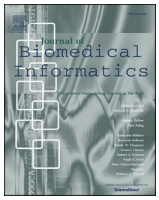
Journal of Biomedical Informatics 149 (2024) 104568

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin



Enhanced family history-based algorithms increase the identification of individuals meeting criteria for genetic testing of hereditary cancer syndromes but would not reduce disparities on their own

Richard L. Bradshaw [a,b,1], Kensaku Kawamoto [a,b,2], Jemar R. Bather [f,g,3], Melody S. Goodman [f,g,6], Wendy K. Kohlmann [b,c,d], Daniel Chavez-Yenter [d,e,5], Molly Volkmar [d], Rachel Monahan [h], Kimberly A. Kaphingst [d,e,4], Guilherme Del Fiol [a,b,7,*]

[a] Department of Biomedical Informatics, University of Utah, Salt Lake City, UT, USA
[b] University of Utah Health, Salt Lake City, UT, USA
[c] Department of Population Health Sciences, University of Utah, Salt Lake City, UT, USA
[d] Huntsman Cancer Institute, University of Utah, Salt Lake City, UT, USA
[e] Department of Communication, University of Utah, Salt Lake City, UT, USA
[f] Department of Biostatistics, New York University School of Global Public Health, New York, NY, USA
[g] Center for Anti-racism, Social Justice, & Public Health, New York University School of Global Public Health, New York, NY, USA
[h] New York University Langone Health, New York, NY, USA

ARTICLE INFO

Keywords:
Electronic health records
Healthcare disparities
Algorithm development
Genetic testing
Hereditary cancer syndromes

ABSTRACT

Objective: This study aimed to 1) investigate algorithm enhancements for identifying patients eligible for genetic testing of hereditary cancer syndromes using family history data from electronic health records (EHRs); and 2) assess their impact on relative differences across sex, race, ethnicity, and language preference.
Materials and Methods: The study used EHR data from a tertiary academic medical center. A baseline rule-base algorithm, relying on structured family history data (structured data; SD), was enhanced using a natural language processing (NLP) component and a relaxed criteria algorithm (partial match [PM]). The identification rates and differences were analyzed considering sex, race, ethnicity, and language preference.
Results: Among 120,007 patients aged 25–60, detection rate differences were found across all groups using the SD (all $P < 0.001$). Both enhancements increased identification rates; NLP led to a 1.9 % increase and the relaxed criteria algorithm (PM) led to an 18.5 % increase (both $P < 0.001$). Combining SD with NLP and PM yielded a 20.4 % increase ($P < 0.001$). Similar increases were observed within subgroups. Relative differences persisted across most categories for the enhanced algorithms, with disproportionately higher identification of patients who are White, Female, non-Hispanic, and whose preferred language is English.
Conclusion: Algorithm enhancements increased identification rates for patients eligible for genetic testing of hereditary cancer syndromes, regardless of sex, race, ethnicity, and language preference. However, differences in identification rates persisted, emphasizing the need for additional strategies to reduce disparities such as addressing underlying biases in EHR family health information and selectively applying algorithm enhancements for disadvantaged populations. Systematic assessment of differences in algorithm performance across population subgroups should be incorporated into algorithm development processes.

* Corresponding author at: 421 Wakara Way, Suite 140, Salt Lake City, UT 84108, USA.
E-mail address: guilherme.delfiol@utah.edu (G. Del Fiol).
[1] Richard L. Bradshaw: 0000-0001-7363-0327.
[2] Kensaku Kawamoto: 0000-0003-4282-9338.
[3] Jemar R. Bather: 0000-0002-0285-3678.
[4] Kimberly A. Kaphingst: 0000-0003-2668-9080.
[5] Daniel Chavez-Yenter: 0000-0001-7764-4443.
[6] Melody S. Goodman: 0000-0001-8932-624X.
[7] Guilherme Del Fiol: 0000-0001-9954-6799.

https://doi.org/10.1016/j.jbi.2023.104568
Received 21 September 2023; Received in revised form 7 November 2023; Accepted 7 December 2023
Available online 9 December 2023

1. Introduction

Recent widespread adoption of electronic health records (EHRs) is revolutionizing healthcare. Larger volumes of electronic healthcare data promote the discovery of new evidence and the delivery of evidence-based healthcare interventions. EHR data-driven algorithms continue to evolve providing clinical insights that assist healthcare professionals to identify patients who benefit from certain healthcare services. However, EHR algorithms that rely on patient data can only assist individuals who have the required data. Populations who experience data poverty (those who have disproportionately incomplete and inaccurate EHR data) are overlooked, exacerbating disparities in healthcare outcomes, particularly for medically underserved and marginalized populations. [1].

Individualized cancer risk evaluation is one area that could benefit from population-based approaches driven by algorithms over EHR data. Early identification of individuals at higher inherited risk for developing cancer is critical for personalized cancer prevention and to reduce disparities in morbidity and mortality, particularly among individuals from historically marginalized groups.[2] For example, the US Preventive Services Task Force recommends genetic testing be incorporated into risk assessment of patients with personal or family history of breast or ovarian cancer in order to identify those individuals with cancer risk levels warranting increased screening or risk reducing surgery.[3] Similarly, the US Multi-Society Task Force on Colorectal Cancer recommends including family history in tailoring colorectal cancer screening.[4] Estimates based on family history indicate that the prevalence of individuals with familial risk is 13 % for breast cancer and 5 % for colorectal cancer.[5] However, despite increased availability and lower cost of genetic testing, the majority of individuals meeting evidence-based criteria for genetic testing of hereditary syndromes have not received genetic services.[6,7].

The Genetic Cancer Risk Detector (GARDE) platform is an EHR innovation that uses algorithms to identify patient populations that meet criteria set by National Comprehensive Cancer Network (NCCN) guidelines for genetic testing of hereditary cancer syndromes using patients' family health history from the EHR.[8,9] GARDE has been used to support the Broadening the Reach, Impact, and Delivery of Genetic Services (BRIDGE) trial. BRIDGE is a randomized controlled trial with 3,073 patients who receive primary care at the University of Utah Health (UHealth) and New York University (NYU) Langone Health. The trial compared two models of patient outreach and education (enhanced standard of care versus automated chatbot) to offer eligible patients access to genetic testing for hereditary breast, ovarian, and colorectal cancer syndromes.[10] In a recent study that analyzed family history data extracted from UHealth and NYU, our group discovered substantial disparities across sex, race, ethnicity, and preferred language in the availability and completeness of family health history documentation and consequently in the identification of NCCN-eligible patients at both organizations.[11].

Given the effect of information presence bias discovered in family history data and GARDE's dependence on structured family history data, the authors formulated two methods to mitigate missing data with the goal of reducing the discovered disparities: extracting family history attributes such as age of disease onset using natural language processing (NLP) over family history comments fields; and relaxing algorithm criteria to identify individuals who partially match criteria. As such, the objective of this study was to investigate these new methods comparing 1) identification rates of eligible patients; and 2) demographic differences according to sex, race, ethnicity, and preferred language.

1.1. Statement of significance

Problem. Computer algorithms over EHR data are promising approaches to identify patients who may benefit from certain healthcare services, such as genetic testing, but have the potential to exacerbate health disparities.

What is already known. A previous study has shown significant disparities in family health history documentation in the EHR in terms of sex, race, ethnicity, and language preference.[11].

What this paper adds. This study investigated EHR algorithms to help address demographic differences in the identification of patients meeting family history-based criteria for genetic testing of hereditary cancer syndromes. The study provides a method that could be used as a part of EHR algorithm development to deliberately assess potential algorithm disparities.

2. Methods

2.1. Setting

The setting for the study was the UHealth system, a tertiary academic medical center and one of the largest healthcare delivery systems in the Intermountain West.

2.2. Study population

Study participants included individuals aged 25 to 60 years who had completed a primary care visit at UHealth within a 2-year time window between September 16, 2020 and September 15, 2022.

2.3. Data

Retrospective EHR data were extracted from UHealth's enterprise data warehouse (EDW). Patient demographics (sex, race, ethnicity, and primary language) and structured family history along with unstructured comments associated with structured family history assertions were extracted for all individuals in the study population.

2.4. Algorithms

GARDE's baseline algorithm uses structured family health history data (structured data; SD) from the EHR, which are stored in three discrete data fields: *disease of interest* (e.g., breast cancer, colorectal cancer), family member *relationship* (e.g., mother, sister, paternal grandfather), and *age of onset* in years. Each NCCN criteria relies at a minimum on *disease of interest* and *relationship* (e.g., first-degree relative with pancreatic cancer), while a subset of the criteria also relies on cancer age of onset (e.g., first- or second-degree relative with breast cancer at age $\leq$ 45 years). Details about the logic and evaluation of this algorithm are available elsewhere.[8,9].

There are important limitations with the family health history SD: 1) *age of onset* in years is often missing, and 2) *disease of interest* and *relationship* codes often lack specificity. Both limitations impact the ability to process NCCN criteria. However, the family history module in the EHR provides a free-text *comments* field adjacent to structured data items to allow users to add unconstrained information such as age of onset using fuzzy terms (e.g., "in her thirties"), a more specific disease of interest (e.g., "breast cancer" when SD only has a code for "cancer"), and the family member's side of the family (e.g., "paternal"). Two enhancements were investigated to address both limitations: 1) a natural language processing (NLP) algorithm to extract information from the *comments* field and 2) a relaxation of the NCCN *age of onset*-specific criteria allowing for partial matches (PM) when age of onset is missing. An overview of each method follows below.

2.4.1. Structured Data + Natural language processing (SD + NLP)

Upon review of the family history statements, common information patterns emerged observing the free text *comments* as follows:

1. A general coded disease of interest (e.g., code for "cancer") is further specified in the comments (e.g., "breast") rather than selecting a pre-coordinated breast cancer code.
2. The side of the family (e.g., "paternal") is specified in the comments to supplement a coded family member relationship (e.g., code for "uncle") rather than selecting a pre-coordinated "paternal uncle" code.
3. Age of onset values and ranges are provided as text rather than an integer (e.g., "in her thirties").

The NLP algorithm augments the SD by extracting information from the free-text *comments* for all three patterns above. Table 1 provides three examples of family history assertions using a combination of SD and data extracted from the free text comments field. The NLP component uses a rule-based approach with 95 % sensitivity and 99 % precision in correctly extracting fragments of family history information (i.e., *disease of interest*, *relationship side of the family*, and *age of onset*) from the comments fields. Details of the development and evaluation of the SD + NLP algorithm are described elsewhere.[12,13].

2.4.2. Partial match (PM)

The PM strategy relaxes 3 of the 11 NCCN criteria rules that depend on *age of onset*. The other 8 rules that do not rely on age of onset are used with no modification. The rules before and after the relaxed approach are as follows:

1. From "First or second degree relative with breast cancer and age of onset < 50" to "First or second degree relative with breast cancer"
2. "First or second degree relative with colon cancer and age of onset < 50" to "First or second degree relative with colon cancer."
3. "First or second degree relative with endometrial cancer and age of onset < 50" to "First or second degree relative with endometrial cancer."

With the partial match algorithm, patients who meet any of the 3 rules above, or any of the 8 rules that do not rely on age of onset, are considered as meeting criteria.

2.5. Hypotheses

Two null hypotheses were tested regarding the identification of patients who potentially meet NCCN criteria for genetic evaluation of hereditary cancer syndromes:

1. Compared to SD, enhanced algorithms do not increase the identification rate for each sex, race, ethnicity, and language.
2. Compared to SD, enhanced algorithms do not reduce differences in the identification rate across sex, race, ethnicity, and language.

Table 1
Structured family history statements before (SD) and after NLP (SD + NLP) for three sample cases. Bolded values indicate NLP-added improvements.

ID	(Code) Method	(Code) Relationship	(Code) Disease of Interest	(Integer) Age of Onset	(Free Text) Comments
1	SD	Mother	Cancer		breast
1	SD + NLP	Mother	**Breast Cancer**		breast
2	SD	Uncle	Prostate Cancer		paternal
2	SD + NLP	**Paternal Uncle**	Prostate Cancer		paternal
3	SD	Maternal Aunt	Breast Cancer		In her thirties
3	SD + NLP	Maternal Aunt	Breast Cancer	**30***	In her thirties

* NLP's raw output was 30–39 and GARDE used the lower bound of the range (30).

2.6. Statistical analysis

For the study population we computed descriptive statistics on patients; sex, race, ethnicity, and primary language. Fisher's exact test and Pearson's Chi-squared test were used to test differences in demographic characteristics between those who met the algorithm criteria under the SD condition and those who did not. We used generalized estimating equations[14] (GEEs) with a binomial variance, an independence correlation structure, and an identity link to estimate the percentage-point differences (absolute proportion changes) between algorithm enhancements. Let Y_{ij} denote the binary response of whether the i^{th} patient met the criteria under the j^{th} algorithm enhancement ($j = 1, 2, 3, 4$), where $j = 1$ is the structured data algorithm condition. The population-averaged model for the proportion of patients meeting the algorithm criteria is

$$\pi_{ij} = \beta_1 + \beta_2 X_{ij2} + \beta_3 X_{ij3} + \beta_4 X_{ij4},$$

where $\pi_{ij} = \mathrm{E}(Y_{ij}|X_{ij})$. The parameter β_1 is the proportion of patients who met the criteria under the SD condition. The parameters β_2 (SD + NLP), β_3 (SD + NLP + PM), and β_4 (SD + PM) represent the percentage-point differences (increase or decrease) when compared to using the SD algorithm only.

Multivariable logistic regression[15] models were used to estimate adjusted odds ratios and 95 % confidence intervals for associations between patient demographics and meeting the criteria at each step. Step 1 included all patients. Step 2 included patients that did not meet the criteria using SD + NLP. Step 3 included patients that did not meet the criteria using NLP. These multivariable models were used to compute predicted probabilities by sex, race, ethnicity, and primary language. We used R[16] to perform all statistical analyses and set statistical significance at 0.05.

3. Results

The analysis included 120,007 patients aged 25 to 60 years who had a primary care visit at UHealth in a 2-year window between September 16th 2020 and September 15th 2022. Of those, 70,666 (58.9 %) were female, 88,974 (74.2 %) identified as White, 96,187 (80.2 %) identified as non-Hispanic/Latino, and 110,026 (91.8 %) had English as their preferred language recorded in the EHR (Table 2). Using the SD algorithm, 5,430 (4.5 %) patients met the criteria for genetic evaluation of hereditary cancer syndromes. Using this algorithm, significant differences in identification rates were found in terms of sex, race, ethnicity, and language preference (all $P < 0.001$).

3.1. Do algorithm enhancements increase the identification rate of eligible patients?

Overall, GEE estimates of percentage-point differences for algorithm enhancements showed that both enhancements led to significant increases in identification rates (Table 3). Compared with SD alone, adding NLP or PM increased the identification rate by 1.9 % and 18.5 %, respectively (both $P < 0.001$). Combining SD with NLP and PM led to the highest increase of 20.4 % ($P < 0.001$). Significant increases in identification rates were found within each sex, race, ethnicity, and language preference category (all $P < 0.001$). Under each algorithm enhancement (SD, SD + NLP, SD + NLP + PM), highest increases were found for females (6.1 %, 8.5 %, and 31.9 %, respectively), patients identified as White (5.1 %, 7.3 %, and 27.8 %), non-Hispanic patients (4.8 %, 6.9 %, and 26.5 %), and those whose preferred language was English (4.8 %, 6.8 %, and 26.2 %).

The NLP algorithm extracted 2,033 *disease of interest* instances, 91 *relationship* instances, and 680 *age of onset* instances. This led to the identification of 2,268 additional patients. The most frequent contributions were due to patients meeting breast cancer criteria due to the

Table 2
Patient characteristics and identification rates using the structured data (SD) algorithm.

		Overall	Met Criteria (SD)		
			No	Yes	
Characteristic	**N**	N = 120,007	N = 114,577	N = 5,430	**P value**[1]
Sex, No. (%)	120,007				**<0.001**
Female		70,666 (58.88 %)	66,349 (57.91 %)	4,317 (79.50 %)	
Male		49,293 (41.08 %)	48,180 (42.05 %)	1,113 (20.50 %)	
Unknown/Did not disclose		48 (0.04 %)	48 (0.04 %)	0 (0.00 %)	
Race, No. (%)	119,937				**<0.001**
White		88,974 (74.18 %)	84,408 (73.71 %)	4,566 (84.10 %)	
American Indian/ Alaska Native		801 (0.67 %)	777 (0.68 %)	24 (0.44 %)	
Asian		5,824 (4.86 %)	5,684 (4.96 %)	140 (2.58 %)	
Black or African American		3,258 (2.72 %)	3,174 (2.77 %)	84 (1.55 %)	
Native Hawaiian/ Pacific Islander		1,878 (1.57 %)	1,827 (1.60 %)	51 (0.94 %)	
Other		16,018 (13.36 %)	15,546 (13.58 %)	472 (8.69 %)	
Unknown/Did not disclose		3,184 (2.65 %)	3,092 (2.70 %)	92 (1.69 %)	
Ethnicity, No. (%)	119,943				**<0.001**
non-Hispanic/ Latino		96,187 (80.19 %)	91,542 (79.94 %)	4,645 (85.57 %)	
Hispanic/Latino		19,226 (16.03 %)	18,577 (16.22 %)	649 (11.96 %)	
Unknown/Did not disclose		4,530 (3.78 %)	4,396 (3.84 %)	134 (2.47 %)	
Primary Language, No. (%)	119,802				**<0.001**
English		110,026 (91.84 %)	104,746 (91.58 %)	5,280 (97.24 %)	
Spanish		6,075 (5.07 %)	5,964 (5.21 %)	111 (2.04 %)	
Other		3,701 (3.09 %)	3,662 (3.20 %)	39 (0.72 %)	

[1] Fisher's exact test; Pearson's Chi-squared test.

following:

- 771 (34 % of the new patients) patients met criteria due to a 1st degree relative with pancreatic cancer
- 641 (28.3 % of the new patients) due to a 1st or 2nd degree relative with ovarian cancer
- 318 (14.1 % of the new patients) due to a 1st or 2nd degree relative with breast cancer and age of onset at less than 45 years old

The relaxed PM algorithm led to 22,212 additional patients meeting criteria based on the three rules that relied on age of onset. Of those:

- 17,088 (76.9 %) additional patients met breast cancer criteria due to a 1st or 2nd degree relative with breast cancer and unknown age of onset
- 12,272 (55.2 %) met colorectal cancer criteria due to a 1st degree relative with colon cancer and unknown age of onset
- and 830 (3.7 %) met colorectal cancer criteria due to a 1st degree relative with endometrial cancer and unknown age of onset

Table 3
Generalized estimating equations estimates of percentage-point differences for algorithm enhancements: A = SD, B = SD + NLP, C = SD + NLP + PM, and D = SD + PM.

					Percentage-point difference		
					NLP effect	NLP + PM effect	PM effect
Sample	*A (%)*	*B (%)*	*C (%)*	*D (%)*	*B-A*	*C-A*	*D-A*
Overall	4.52	6.41	24.92	23.03	1.89	20.40	18.51
Sex							
Female	6.11	8.48	31.85	29.48	2.37	25.74	23.37
Male	2.26	3.46	15.02	13.82	1.21	12.76	11.56
Race							
White	5.13	7.30	27.81	25.64	2.17	22.68	20.51
American Indian/Alaska Native	3.00	4.49	17.98	16.48	1.50	14.98	13.48
Asian	2.40	3.30	15.95	15.06	0.89	13.55	12.65
Black or African American	2.58	3.41	13.01	12.19	0.83	10.44	9.61
Native Hawaiian/ Pacific Islander	2.72	4.26	17.57	16.03	1.54	14.86	13.31
Other	2.95	4.05	17.41	16.31	1.10	14.46	13.37
Unknown/Did not disclose	2.89	4.24	16.90	15.55	1.35	14.01	12.66
Ethnicity							
non-Hispanic	4.83	6.89	26.47	24.40	2.06	21.64	19.57
Hispanic	3.38	4.56	19.07	17.89	1.18	15.69	14.51
Unknown/Did not disclose	2.96	4.19	17.15	15.92	1.24	14.19	12.96
Primary Language							
English	4.80	6.81	26.15	24.14	2.01	21.36	19.34
Spanish	1.83	2.45	13.14	12.51	0.63	11.31	10.68
Other	1.05	1.46	8.97	8.57	0.41	7.92	7.51

Note: P values for all tested differences were < 0.001; Models for Unknow/Did not disclose Sex did not converge due to sample size; The following comparisons are not included due to redundancy: D-C = -(B-A); C-B = D-A.

3.2. Do algorithm enhancements affect relative differences across groups?

Overall, multivariable logistic regression showed that relative differences in identification rates persisted with each algorithm enhancement across most categories (Table 4). Compared to females, males had lower identification odds using the SD (OR = 0.35, 95 % CI: 0.33, 0.38), SD + NLP (OR = 0.48, 95 % CI: 0.43, 0.52), and PM (OR = 0.39, 95 % CI: 0.38, 0.40) algorithms.

Compared to patients identified as White, lower identification odds were found using the SD, NLP, and PM algorithms for American Indian/ Alaska Native (ORs = 0.53, 0.65, and 0.54 respectively; all P < 0.001), Asian (ORs = 0.52, 0.45, and 0.57; all P < 0.001), Black or African American (ORs = 0.60, 0.45, 0.45; all P < 0.001), and Native Hawaiian/ Pacific Islander (OR = 0.53, 0.70, 0.57; all P < 0.001).

Compared to patients who identified as non-Hispanic, Hispanic patients had lower identification odds using the SD (OR = 0.90; P = 0.076),

Table 4
Multivariable logistic regression models showing predictors of meeting the criteria at each algorithm enhancement: Step 1 population = all patients analyzed with the SD algorithm; Step 2 population = patients who did not meet the criteria using the SD algorithm, analyzed with SD + NLP algorithm; Step 3 population = patients who did not meet the criteria using SD or SD + NLP, analyzed using PM algorithm.

	Step 1 (SD)			**Step 2 (SD + NLP)**			**Step 3 (PM)**		
	N = 119,677			N = 114,249			N = 111,982		
Characteristic	**OR**	**95 % CI**	**P value**	**OR**	**95 % CI**	**P value**	**OR**	**95 % CI**	**P value**
Sex									
Female	—	—		—	—		—	—	
Male	0.35	0.33, 0.38	**<0.001**	0.48	0.43, 0.52	**<0.001**	0.39	0.38, 0.40	**<0.001**
Race									
White	—	—		—	—		—	—	
American Indian/Alaska Native	0.53	0.35, 0.79	**0.003**	0.65	0.35, 1.10	0.14	0.54	0.44, 0.66	**<0.001**
Asian	0.52	0.43, 0.61	**<0.001**	0.45	0.33, 0.58	**<0.001**	0.57	0.53, 0.62	**<0.001**
Black or African American	0.60	0.48, 0.75	**<0.001**	0.45	0.30, 0.65	**<0.001**	0.45	0.40, 0.51	**<0.001**
Native Hawaiian/Pacific Islander	0.53	0.39, 0.69	**<0.001**	0.70	0.47, 1.00	0.06	0.57	0.49, 0.65	**<0.001**
Other	0.74	0.65, 0.83	**<0.001**	0.76	0.62, 0.92	**0.006**	0.74	0.70, 0.79	**<0.001**
Unknown/Did not disclose	0.73	0.55, 0.97	**0.034**	0.93	0.60, 1.41	0.70	0.73	0.63, 0.85	**<0.001**
Ethnicity									
non-Hispanic/Latino	—	—		—	—		—	—	
Hispanic/Latino	0.90	0.81, 1.01	0.076	0.73	0.61, 0.87	**<0.001**	0.83	0.78, 0.88	**<0.001**
Unknown/Did not disclose	0.79	0.62, 0.99	**0.049**	0.66	0.45, 0.94	**0.028**	0.75	0.66, 0.85	**<0.001**
Language									
English	—	—		—	—		—	—	
Spanish	0.46	0.37, 0.57	**<0.001**	0.44	0.30, 0.61	**<0.001**	0.61	0.56, 0.67	**<0.001**
Other	0.28	0.20, 0.38	**<0.001**	0.27	0.16, 0.44	**<0.001**	0.41	0.36, 0.47	**<0.001**

OR = Odds Ratio, CI = Confidence Interval.

NLP (OR = 0.73; P < 0.001), and PM (OR = 0.83; P < 0.001). Compared with patients with an English preference, patients who preferred Spanish had lower identification odds with the SD, SD + NLP, and PM algorithms (OR = 0.46, 0.44, and 0.61 respectively; P < 0.001).

Table 5 shows that patients identified as non-Hispanic White females who prefer English had the highest probability of meeting algorithm criteria with the SD (0.071; CI: 0.069, 0.073), SD + NLP (0.030; CI: 0.029, 0.032), and PM (0.292; CI: 0.288, 0.296) algorithms. These probabilities were considerably higher than among other groups, the lowest of which ranged from 0.003 (CI: 0.002, 0.005) using the SD algorithm to 0.022 (0.018, 0.027) using the PM algorithm (Table 4).

Table 5
Top and bottom five predicted probabilities of meeting genetic testing criteria based on Sex, Race, Ethnicity and Language comparing 3 stepped populations. Step 1 - all patients. Step 2 - patients that did NOT meet SD + NLP criteria. Step 3 - patients that did NOT meet NLP criteria.

Rank	*Sex*	*Race*	*Ethnicity*	*Language*	*Probability*	*95 % CI*
Step 1 population = all patients (N = 119,677) – SD algorithm						
Top 5	Female	White	non-Hispanic/Latino	English	0.071	(0.069, 0.073)
	Female	White	Hispanic/Latino	English	0.065	(0.058, 0.072)
	Female	White	Unknown/Did not disclose	English	0.057	(0.045, 0.071)
	Female	Other	non-Hispanic/Latino	English	0.053	(0.047, 0.060)
	Female	Unknown/Did not disclose	non-Hispanic/Latino	English	0.053	(0.040, 0.069)
Bottom 5	Male	Native Hawaiian/Pacific Islander	Hispanic/Latino	Other	0.004	(0.002, 0.006)
	Male	Asian	Hispanic/Latino	Other	0.003	(0.002, 0.005)
	Male	American Indian/Alaska Native	Unknown/Did not disclose	Other	0.003	(0.002, 0.006)
	Male	Native Hawaiian/Pacific Islander	Unknown/Did not disclose	Other	0.003	(0.002, 0.005)
	Male	Asian	Unknown/Did not disclose	Other	0.003	(0.002, 0.005)
Step 2 population = patients that did not meet the criteria using SD (N = 114,249) – SD + NLP algorithm						
Top 5	Female	White	non-Hispanic/Latino	English	0.030	(0.029, 0.032)
	Female	Unknown/Did not disclose	non-Hispanic/Latino	English	0.028	(0.019, 0.042)
	Female	Other	non-Hispanic/Latino	English	0.023	(0.019, 0.028)
	Female	White	Hispanic/Latino	English	0.022	(0.019, 0.027)
	Female	Native Hawaiian/Pacific Islander	non-Hispanic/Latino	English	0.021	(0.015, 0.031)
Bottom 5	Male	American Indian/Alaska Native	Unknown/Did not disclose	Other	0.002	(0.001, 0.004)
	Male	Black or African American	Hispanic/Latino	Other	0.001	(0.001, 0.003)
	Male	Asian	Hispanic/Latino	Other	0.001	(0.001, 0.002)
	Male	Black or African American	Unknown/Did not disclose	Other	0.001	(0.001, 0.002)
	Male	Asian	Unknown/Did not disclose	Other	0.001	(0.001, 0.002)
Step 3 population = patients that did not meet the criteria using NLP (N = 111,982) - PM						
Top 5	Female	White	non-Hispanic/Latino	English	0.292	(0.288, 0.296)
	Female	White	Hispanic/Latino	English	0.255	(0.244, 0.266)
	Female	White	Unknown/Did not disclose	English	0.236	(0.215, 0.259)
	Female	Other	non-Hispanic/Latino	English	0.235	(0.223, 0.247)
	Female	Unknown/Did not disclose	non-Hispanic/Latino	English	0.232	(0.207, 0.259)
Bottom 5	Male	Asian	Unknown/Did not disclose	Other	0.028	(0.023, 0.033)
	Male	Native Hawaiian/Pacific Islander	Unknown/Did not disclose	Other	0.027	(0.022, 0.034)
	Male	American Indian/Alaska Native	Unknown/Did not disclose	Other	0.026	(0.020, 0.034)
	Male	Black or African American	Hispanic/Latino	Other	0.024	(0.020, 0.029)
	Male	Black or African American	Unknown/Did not disclose	Other	0.022	(0.018, 0.027)

4. Discussion

Two algorithm enhancements using NLP (SD + NLP) and a relaxed eligibility criteria (PM) were investigated to increase the identification rate and reduce demographic differences among patients meeting evidence-based criteria for genetic testing of hereditary cancer syndromes. Both enhancements incrementally increased the identification rates across all groups according to sex, race, ethnicity, and language preference significantly. Therefore, the enhanced algorithms have the potential to benefit a substantially larger number of patients. Yet, neither enhancement substantially decreased the systematic relative differences that were discovered in the output of the SD algorithm.[11].

4.1. Enhanced algorithms

This study demonstrated algorithm enhancements testing two approaches to addressing missing data that led to significant increases in the identification of patients who would benefit from genetic testing of hereditary cancer syndromes. Other healthcare use cases could benefit from similar approaches to address missing data. Placing free-text fields adjacent to structured data fields is a common pattern used to supplement structured data in EHRs. Using NLP to extract information from these fields could be valuable in other similar use cases. For example, although the majority of data attributes in drug prescriptions (e.g., drug product, strength, form, route) are collected in structured format, provider directions for taking the medication (e.g., "Take one tablet by mouth every four to six hours") are often stored in unstructured format and are critical for use cases where it is necessary to compute the total dose of medication taken over time.[17].

Surprisingly, the information extracted by NLP that had the highest yield in additional patients meeting testing criteria was for *disease of interest*. Users often selected a generic code (i.e., code for "cancer") from a drop-down list and typed the specific cancer (e.g., "pancreatic") in the free-text comments field, even though a comprehensive list of cancer types is provided as a drop-down list in the family history section of the EHR. This finding suggests that improved usability may help users select specific cancer codes for the *disease of interest* more efficiently. Another type of information that resulted in the identification of additional patients was *age of onset*. Although a discrete field is available in the EHR for users to type the age of onset as an integer, users most often prefer to enter a fuzzy range as text (e.g., low 30 s), possibly because patients may not know the exact age of onset. Possible improvements include the provision of a drop-down list with relevant ranges. In fact, such an improvement has recently been implemented by the Epic EHR, which now allows the documentation of *age of onset* as a range in addition to an integer. In addition to user interface improvements, another potential approach is to apply NLP methods to extract family history information from clinical notes, since specific providers may not be using the EHR's dedicated family history section to document family history.

The PM approach helped identify a large number of additional individuals who did not meet full criteria, due to missing age of onset. Targeted approaches could be designed to complete missing data for these patients and confirm their eligibility for genetic testing. For example, a population-based outreach approach could be used to target patients with documented family history of breast, endometrial, and colorectal cancer, but no age of onset specified. Digital tools, such as patient portals or chatbots, could be used to ask one simple question to confirm whether the approximate age of onset was above or below a threshold value. Other approaches could be used to prioritize patients for additional data collection and outreach, such as data imputation and machine learning, with the goal of identifying patients with the highest likelihood of meeting criteria. Therefore, PM algorithms are a promising approach to identify targeted patient cohorts that may benefit from additional data collection.

4.2. Differences by demographic characteristics

The COVID-19 pandemic recently magnified the need to disaggregate healthcare data by demographic characteristics when it became clear that specific groups were being disproportionally affected.[18] A study examining demographic disparities in algorithmic performance for detecting in-hospital patient deterioration has found analogous discrepancies post-hoc, years after the algorithm's widespread implementation, and by researchers not initially involved in its development.[19] This kind of validation is rarely done as a part of algorithm development, before algorithms are widely implemented in clinical settings. The present study illustrates a method that could be used as a part of algorithm development to systematically quantify and compare differences by demographic characteristics resulting from multiple EHR data-driven algorithms. Findings from such assessments could be used to further guide algorithm development and to design other mitigation strategies, such as improved datasets.

In the present and other similar use cases, it is possible that certain kinds of algorithm-induced disparities cannot be addressed through algorithmic approaches. Rather, directed strategies are needed to address systemic biases in the underlying data. For example, targeted data collection approaches could be implemented such as using community health workers and self-administered questionnaires to collect abbreviated family history,[20] digital health navigators to improve access to patient portals for pre-visit questionnaires,[21] and proactive patient outreach via digital tools to complete missing attributes in family history records.

There is movement suggesting all women be screened for BRCA mutations[22], potentially addressing genetic testing inequities for women. However, national guidelines such as NCCN[23,24] and the US Preventive Services Task Force[3] (USPSTF) still recommend a risk-based approach based on family history. Nevertheless, the GARDE architecture allows algorithms to be updated over time to reflect changes in national recommendations for genetic testing.

5. Limitations

This study had several limitations. First, the analysis assumes that the prevalence of hereditary cancer syndromes is similar among different groups, when there may be differences attributable to genetic and socio-environmental factors or even bias in the underlying NCCN criteria. Yet, it is unlikely that differences in prevalence among groups would be as substantial as those found in this study. Second, although the PM approach led to a substantial increase in the identification rate, only an unknown subset of those patients will actually meet criteria after additional steps are taken to collect missing data. Still, as described above, the PM algorithm can be used to identify cohorts for targeted data collection approaches. Last, this study was conducted at one academic healthcare system that uses a specific EHR product and provides care to a population in one US state (Utah) which does not reflect the national distributions of race and ethnicity. Thus, it is unknown if the results are generalizable to other settings. Still, the proposed algorithm enhancement approaches and disparity analysis method are healthcare system and EHR-agnostic, and therefore could be adapted or adopted in other settings.

6. Conclusion

This study found substantial relative differences in the algorithmic identification of individuals meeting family history-based criteria for genetic testing of hereditary cancer syndromes. Although algorithm enhancements increased the overall identification rate, relative differences across groups persisted. Directed approaches (e.g., self-administered questionnaires to collect abbreviated family history, digital tools) are needed to address underlying differences in EHR data availability and completeness by demographic group as well as

potentially the prioritized application of algorithm enhancements to populations with missing data. Algorithm development studies should systematically and proactively assess disparities in algorithm performance as a part of algorithm development.

Funding

This study was supported by grants U01CA232826S1, U24CA204800, and 1U24CA274582 from the National Cancer Institute (NCI) of the National Institutes of Health (NIH).

CRediT authorship contribution statement

Richard L. Bradshaw: Supervision, Writing – original draft, Methodology, Software, Data curation. **Kensaku Kawamoto:** Writing – review & editing. **Jemar R. Bather:** Methodology, Writing – review & editing. **Melody S. Goodman:** Methodology, Writing – review & editing. **Wendy K. Kohlmann:** Writing – review & editing. **Daniel Chavez-Yenter:** Writing – review & editing. **Molly Volkmar:** Project administration. **Rachel Monahan:** Project administration. **Kimberly A. Kaphingst:** Supervision, Funding acquisition, Writing – review & editing. **Guilherme Del Fiol:** Conceptualization, Supervision, Funding acquisition, Writing – original draft.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: [Guilherme Del Fiol reports financial support was provided by National Institutes of Health. Guilherme Del Fiol is an editorial board member of the Journal of Biomedical Informatics. Kensaku Kawamoto has received consulting honoraries from Pfizer, RTI International, University of California at San Francisco, Indiana University, Korean Society of Medical Informatics, NORC at University of Chicago, Regenstrief Foundation, University of Pennsylvania, Yale University, and Security Risk Solutions].

References

[1] H. Ibrahim, X. Liu, N. Zariffa, A.D. Morris, A.K. Denniston, Health data poverty: an assailable barrier to equitable digital health care, Lancet Digit Health. 3 (4) (2021) e260–e265, https://doi.org/10.1016/S2589-7500(20)30317-4.

[2] R.G. Caffrey, Advocating for equitable management of hereditary cancer syndromes, J Genet Couns. 31 (3) (2022) 584–589, https://doi.org/10.1002/jgc4.1548.

[3] U.S.P.S.T. Force, D.K. Owens, K.W. Davidson, et al., Risk assessment, genetic counseling, and genetic testing for BRCA-related cancer: US preventive services task force recommendation statement, JAMA 322 (7) (2019) 652–665, https://doi.org/10.1001/jama.2019.10987.

[4] D.K. Rex, C.R. Boland, J.A. Dominitz, et al., Colorectal cancer screening: Recommendations for physicians and patients from the U.S. multi-society task force on colorectal cancer, Gastroenterology 153 (1) (2017) 307–323, https://doi.org/10.1053/j.gastro.2017.05.013.

[5] Scheuner MT, McNeel TS, Freedman AN. Population prevalence of familial cancer and common hereditary cancer syndromes. The 2005 California Health Interview Survey. *Genetics in medicine : official journal of the American College of Medical Genetics*. Nov 2010;12(11):726-310.1097/GIM.0b013e3181f30e9e.

[6] A.W. Kurian, P. Abrahamse, A. Furgal, et al., Germline Genetic Testing After Cancer Diagnosis, JAMA 330 (1) (2023) 43–51, https://doi.org/10.1001/jama.2023.9526.

[7] A.W. Kurian, K.C. Ward, P. Abrahamse, et al., Time Trends in Receipt of Germline Genetic Testing and Results for Women Diagnosed With Breast Cancer or Ovarian Cancer, 2012–2019, J Clin Oncol. 39 (15) (2021) 1631–1640, https://doi.org/10.1200/JCO.20.02785.

[8] G. Del Fiol, W. Kohlmann, R.L. Bradshaw, et al., Standards-based clinical decision support platform to manage patients who meet guideline-based criteria for genetic evaluation of familial cancer, JCO Clin Cancer Inform. 4 (2020) 1–9, https://doi.org/10.1200/CCI.19.00120.

[9] R.L. Bradshaw, K. Kawamoto, K.A. Kaphingst, et al., GARDE: a standards-based clinical decision support platform for identifying population health management cohorts, J Am Med Inform Assoc. 29 (5) (2022) 928–936, https://doi.org/10.1093/jamia/ocac028.

[10] K.A. Kaphingst, W. Kohlmann, R.L. Chambers, et al., Comparing models of delivery for cancer genetics services among patients receiving primary care who meet criteria for genetic evaluation in two healthcare systems: BRIDGE randomized controlled trial, BMC Health Serv Res. 21 (1) (2021) 542, https://doi.org/10.1186/s12913-021-06489-y.

[11] D. Chavez-Yenter, M.S. Goodman, Y. Chen, et al., Association of disparities in family history and family cancer history in the electronic health record with sex, race, hispanic or latino ethnicity, and language preference in 2 large US health care systems, JAMA Netw Open. 5 (10) (2022), https://doi.org/10.1001/jamanetworkopen.2022.34574.

[12] Mowery DL, Kawamoto K, Bradshaw R, et al. Determining Onset for Familial Breast and Colorectal Cancer from Family History Comments in the Electronic Health Record. 2019.

[13] J. Shi, K.L. Morgan, R.L. Bradshaw, et al., Identifying patients who meet criteria for genetic testing of hereditary cancers based on structured and unstructured family health history data in the electronic health record: natural language processing approach, JMIR Med Inform. 10 (8) (2022), https://doi.org/10.2196/37842.

[14] K.-Y. Liang, S.L. Zeger, Longitudinal data analysis using generalized linear models, Biometrika 73 (1) (1986) 13–22, https://doi.org/10.1093/biomet/73.1.13.

[15] B. Hidalgo, M. Goodman, Multivariate or multivariable regression? Am J Public Health. 103 (1) (2013) 39–40, https://doi.org/10.2105/AJPH.2012.300897.

[16] Team. RC. A Language and Environment for Statistical Computing. *Computing*. 08/22/2023 2006;1.

[17] D.R. Harris, D.W. Henderson, A. Corbeau, sig2db: a workflow for processing natural language from prescription instructions for clinical data warehouses, AMIA Jt Summits Transl Sci Proc. 2020 (2020) 221–230.

[18] W. Mude, V.M. Oguoma, T. Nyanhanda, L. Mwanri, C. Njue, Racial disparities in COVID-19 pandemic cases, hospitalisations, and deaths: A systematic review and meta-analysis, J Glob Health. 11 (2021) 05015, https://doi.org/10.7189/jogh.11.05015.

[19] T.F.t. Byrd, B. Southwell, A. Ravishankar, et al., Validation of a proprietary deterioration index model and performance in hospitalized adults, JAMA Netw Open. 6 (7) (2023) e2324176, https://doi.org/10.1001/jamanetworkopen.2023.24176.

[20] L. Marsh, M. Mendoza, Z. Tatsugawa, et al., A community health worker model to support hereditary cancer risk assessment and genetic Testing, Obstet Gynecol. (2023), https://doi.org/10.1097/aog.0000000000005292.

[21] H. Wisniewski, T. Gorrindo, N. Rauseo-Ricupero, D. Hilty, J. Torous, The role of digital navigators in promoting clinical care and technology integration into practice, *Digit Biomark*. Winter 4 (Suppl 1) (2020) 119–135, https://doi.org/10.1159/000510144.

[22] M.C. King, E. Levy-Lahad, A. Lahad, Population-based screening for BRCA1 and BRCA2: 2014 lasker award, JAMA 312 (11) (2014) 1091–1092, https://doi.org/10.1001/jama.2014.12483.

[23] NCCN.org. Genetic/Familial High-Risk Assessment: Breast and Ovarian. October 30, 20Accessed August 1, 20 https://www.nccn.org/guidelines.

[24] NCCN.org. Genetic/Familial High-Risk Assessment: Colorectal. October 30, 2023. Accessed August 1, 2023. https://www.nccn.org/professionals/physician_gls/pdf/genetics_colon.pdf.