Diversity in Machine Learning: A Systematic Review of Text-Based Diagnostic Applications

Lane Fitzsimmons[1] Maya Dewan[2,3] Judith W. Dexheimer[3,4]

[1] College of Agriculture and Life Science, Cornell University, Ithaca, New York, United States
[2] Division of Critical Care Medicine, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio, United States
[3] Department of Pediatrics, University of Cincinnati College of Medicine, Cincinnati, Ohio, United States
[4] Division of Emergency Medicine; Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio, United States

Address for correspondence Lane Fitzsimmons, Cincinnati Children's Hospital Medical Center, 3333 Burnet Avenue, Cincinnati, OH 45229-3026, United States (e-mail: laf229@cornell.edu).

Appl Clin Inform 2022;13:569–582.

Abstract

Objective As the storage of clinical data has transitioned into electronic formats, medical informatics has become increasingly relevant in providing diagnostic aid. The purpose of this review is to evaluate machine learning models that use text data for diagnosis and to assess the diversity of the included study populations.

Methods We conducted a systematic literature review on three public databases. Two authors reviewed every abstract for inclusion. Articles were included if they used or developed machine learning algorithms to aid in diagnosis. Articles focusing on imaging informatics were excluded.

Results From 2,260 identified papers, we included 78. Of the machine learning models used, neural networks were relied upon most frequently (44.9%). Studies had a median population of 661.5 patients, and diseases and disorders of 10 different body systems were studied. Of the 35.9% ($N = 28$) of papers that included race data, 57.1% ($N = 16$) of study populations were majority White, 14.3% were majority Asian, and 7.1% were majority Black. In 75% ($N = 21$) of papers, White was the largest racial group represented. Of the papers included, 43.6% ($N = 34$) included the sex ratio of the patient population.

Discussion With the power to build robust algorithms supported by massive quantities of clinical data, machine learning is shaping the future of diagnostics. Limitations of the underlying data create potential biases, especially if patient demographics are unknown or not included in the training.

Conclusion As the movement toward clinical reliance on machine learning accelerates, both recording demographic information and using diverse training sets should be emphasized. Extrapolating algorithms to demographics beyond the original study population leaves large gaps for potential biases.

Keywords

- ► machine learning
- ► diagnosis
- ► computer assisted
- ► clinical decision-making
- ► electronic health records
- ► gender data

received
November 9, 2021
accepted
April 4, 2022

Georg Thieme Verlag KG,
Rüdigerstraße 14,
70469 Stuttgart, Germany

DOI https://doi.org/10.1055/s-0042-1749119.
ISSN 1869-0327.

Background and Significance

The health care industry produced 2.3 trillion gigabytes of patient data in 2020.[1] Computational systems engineered to solve problems and expose trends create the potential to advance and expedite the completion of tasks in nearly every domain. Informatics has the potential to wield power beyond reducing workload with predetermined instructions. Machine learning algorithms can process data beyond human capacity, automatically improving themselves with the addition of new data. Through machine learning, these algorithms internalize patterns that might have otherwise never been noticed and remained unutilized.

This analysis of large quantities of data is especially applicable to the health care industry. The increasing relevance of informatics in medicine is reflected by the recent and near-total adoption of electronic health records (EHRs), from 9.4% in 2008 to 83.8% in 2015.[2] With computerization has come a wealth of available data for analysis. Both structured and unstructured data hold powerful information, with an estimated 80% of health record information in the unstructured form.[3]

One of the most promising applications of machine learning is to aid in the process of making a diagnosis. With the almost universal reliance on EHRs, hospitals are now able to efficiently collect and store comprehensive patient profiles composed of symptoms, vital signs, family history, demographic information, medications, lab results, and more. Though underlying methodologies of different models vary greatly, machine learning can leverage these massive quantities of patient data to recognize common features of impacted patients. By recognizing trends indicative of a particular condition, machine learning can be used to develop standardized and comprehensive tools to estimate the likelihood that a disease or condition is present.[4–6] In many cases, the varied presentation of diseases in patients and the lack of comprehensive diagnostic parameters make diagnosis “more of an art than a science”.[7] With over 70,000 diagnosis codes for providers to choose from in the International Classification of Diseases tool (ICD-10), the sheer quantity of information warrants automated assistance. The potential for machine learning models to recognize clinically significant patterns and provide data-supported diagnostic recommendations is promising.

Before these algorithms can be widely implemented, however, it is important to note the implications of this automated optimization. Algorithms prioritize the highest predictive accuracy overall, adapting for the most accurate prediction in the majority group.[8] In a diagnostic context, underrepresenting groups in training studies can inhibit the success of the diagnostic tools on these populations making diversity in the study population necessary for algorithmic equity.

Objective

The purpose of this review is to evaluate the literature on machine learning models that use text data to make diagnoses and to assess the diversity of the study population.

Methods

Literature Search

An electronic literature search was performed to gather all papers eligible for inclusion. Three electronic literature databases were utilized: PubMed (MEDLINE), OVID CINAHL, and ISI Web of Science.[9–11] All search terms were defined as Medical Subject Headings (MeSH) in PubMed and as keywords in OVID CINAHL and ISI Web of Science. All MeSH term searches included result-related terminology, such as singular root words. In OVID CINAHL and ISI Web of Science, asterisks were used to search via the word stem. All search timelines began at database instantiation. PubMed search results included results available through July 7th 2020 and OVID CINAHL and ISI Web of Science through July 13th, 2020. Search results included papers only if they contained terms in both of the two necessary concept groupings, machine learning, and diagnosis. We excluded papers with image-based analysis. Papers including terms (1) AND (2) but NOT (3) were eligible for review.

1. Machine learning OR related terms: neural networks, natural language processing, OR knowledge bases.
2. Diagnosis, computer assisted OR clinical decision-making.
3. Diagnostic imaging OR computer-assisted image processing.

Review of Identified Studies

Only papers using machine learning for diagnosis were included. Any models constructed to identify patients with a new illness or problem were considered diagnostic. Qualifying studies also relied entirely or predominantly on text-based data. Papers that analyzed text-based reports, even if referring to image content like radiological notes, were included. Papers that included nontext aspects as one component of a larger analysis that was overall text based were also eligible for inclusion. For example, a paper that included electrocardiogram analysis would not be excluded if it also included a significant analysis of other features that were text based.

Papers were excluded if they relied upon data that are not readily available in standard EHRs. Papers that focused on the specialized analysis of any nontext-based component were also excluded. Components of this nature included electrocardiogram, electroencephalogram, pathology, genomic, or any image-based analysis. Papers that predicted disease progression or anticipated the success of a treatment were also excluded. For example, papers that predicted the severity of disease symptoms or provided recommendations based on medication were excluded. Papers that used animal models or were written in languages other than English were excluded. Papers that provided an overview of the topic but did not apply a model to a clinical dataset were reviewed for additional references but not included. Inclusion and exclusion criteria are listed in ►Table 1.

From the papers identified, the titles and abstracts of each were extracted for review. Two independent reviewers (L.F. and J.W.D) assessed each article for inclusion. Disagreements

Table 1 Inclusion and exclusion criteria for literature articles

Inclusion criteria	Exclusion criteria
• Implemented machine learning • Diagnostic algorithm • Utilized text-based data • English language	• Focused on electrocardiogram, electroencephalogram, pathology, or genomics • Analyzed images • Used animal models • Predicted disease progression • Exclusively reviewed other articles

were resolved by a third reviewer (M.D.). Full text papers were discussed, and inclusion was resolved by consensus.

Data Collection

One reviewer extracted the following data from each included paper: study year, location, disease studied, number of patients, sex ratio of patients, patient race, type of trial, type of text analyzed, data source, algorithms used, type of validation test, performance measures, and primary and secondary outcomes. For papers where the data were obtained from a different location than the study took place, the location was recorded as the location of study, not the data source. If multiple institutions were cited, the primary institution was recorded. It was also noted if each of the studies was completed at an academic medical center and if the disease studied was sex specific. We reached out to authors of papers that lacked demographic information to fill in any gaps. If no response was received after 2 weeks, a follow-up email was sent. If available, an email request was also sent to an alternate author or address.

Analysis

Data were grouped into bins to analyze and present. Studies were grouped by country, year, and the disease studied. Diseases were grouped by the body system they most impacted, and categorizations were reviewed by a physician (MD). Racial groups Caucasian and African American are included in “White” and “Black” groupings, respectively. For each paper, the four racial groups with the highest frequency were listed, and papers that listed additional groups were specified.

Cohen’s kappa statistic was calculated to assess agreement among reviewers. The sample population size was calculated using median and interquartile scores. In many studies, data for race and gender were limited. If numerical values in these categories were not reported, incomplete data such as qualitative phrases, information estimated by authors, or data from only part of the study population were also included.

Results

Review of Identified Studies

A total of 2,260 papers were obtained from the literature keyword search (►Fig. 1). The PubMed search contributed 1,208, CINAHL 673, and ISI Web of Science 379 papers. After removal of duplicates and papers that did not meet the criteria, 78 studies were included. Cohen’s kappa value was 0.26, and there was 91.3% agreement across reviewers.[12,13] ►Table 2 includes the characteristics of all included studies.[14–91]

►Fig. 2 displays the number of studies grouped by publication year. From 1991 through 2014, no more than three studies were published in any year. The total number of published studies is highest in the years 2018 and 2019, with 11 and 17 studies published in each year, respectively. The number of studies per year was displayed only through 2019, as we stopped collecting studies only partway through 2020.

Data Collection

Study Location

Papers from a total of 21 different countries were included in the review. A total of 37 (47.4%) of included studies were from the United States. The countries with the next highest quantity of studies published were China 5 (6.4%) and the United Kingdom 4 (5.1%).

Body Systems

Diseases in a total of 10 body systems were studied. Six papers (7.6%) studied diseases that impacted multiple body systems. The body system that was included in the highest quantity of papers was circulatory, with 19 papers (24.3%). ►Fig. 3 illustrates the body systems studied. Diseases that were studied in multiple papers are listed in ►Table 3.

Artificial Intelligence

Neural networks were the most frequently relied upon algorithm type, with use by 35 papers (44.9%). Of the papers that used neural networks, 19 (54.3%) used backpropagation. Six papers (17.1%) used multilayer perceptron neural networks, five used recurrent neural networks (14.3%), and three used Bayesian neural networks (8.5%).

Logistic regression was used by 19 papers (24.4%), support vector machine by 12 (15.3%), decision tress by 11 (14.1%), and natural language processing by 10 (12.8%). Six papers used Bayesian algorithms (7.7%), with Naïve Bayes used most frequently.

Race

Race data were initially obtained from the original publication or the referenced publicly available dataset. When these data were not available, authors were contacted via email to obtain data on the race of the study population. ►Fig. 4 displays the number of papers with race data available before and after author contact. In total, race data of some form were available from a total of 28 papers (35.9%), including six author estimations based on memory or regional location data. Of 15 authors that responded but did not provide race data, five specified that they no longer had the data available to them (35.7%), seven noted that they never obtained it in the first place (50%), and one did not have permission to share the data (7.1%).

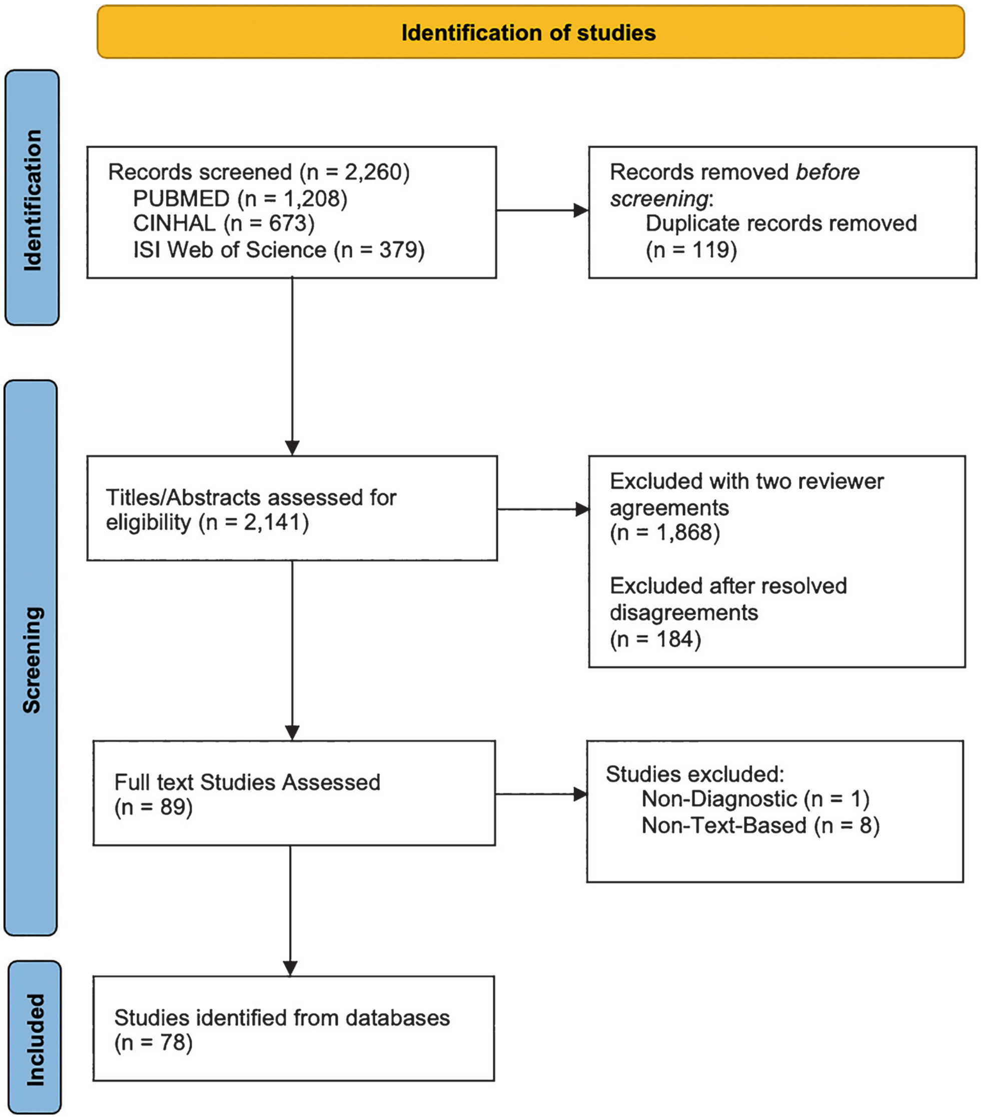
Identification of studies
Identification
Screening
Included
Records screened (n = 2,260)
PUBMED (n = 1,208)
CINHAL (n = 673)
ISI Web of Science (n = 379)
Records removed *before screening*:
Duplicate records removed
(n = 119)
Titles/Abstracts assessed for eligibility (n = 2,141)
Excluded with two reviewer agreements
(n = 1,868)
Excluded after resolved disagreements
(n = 184)
Full text Studies Assessed
(n = 89)
Studies excluded:
Non-Diagnostic (n = 1)
Non-Text-Based (n = 8)
Studies identified from databases
(n = 78)

Fig. 1 Flow diagram of included and excluded studies.

Of the 28 papers with race data available, 16 (57.1%) had patient populations that were predominantly or entirely White or Caucasian. Additionally, five papers listed a white or Caucasian percentage as less than half but still greater than the percentage for any other single group. One paper provided data for only two categories: Black/African American and Hispanic/Latino. Twenty-one papers (75%) included the highest percentage of their study group as White or Caucasian patients. Four study populations were predominantly or entirely Asian, two were predominantly Black or African American, and one was predominantly Pacific people. One study had a “high proportion” of Hispanic patients, but no further information was available. On average, study populations included roughly 13% Black patients and less than 6% Hispanic or Latino patients.

Sex and Gender

Sex and gender information was obtained from the paper or accessible dataset when available. When the data were not available, requests were included in the emails that were sent for race data information. ►Fig. 5 displays the availability of gender data before and after contact. Three papers were for female-specific diseases: breast and ovarian cancers, ectopic pregnancies, and cesarian deliveries. These study populations included exclusively women. Additionally, 90% of Lupus patients are women, so an author studying this disease estimated that her study population was roughly consistent with this ratio.

Of the 45 papers that included sex data for gender-neutral diseases, 45.5% of study patients were women. ►Table 4 illustrates the percent of papers that included less than women in their study populations.

Sample Size

Patients with median of 661.5 and interquartile range of 1,945 patients were included in the papers. This number was taken as the total number of patients, regardless of what percentage of the data were used for training, testing, and

Table 2 Characteristics of included studies

Reference	Publication year	Author	Institution	Disease or condition	Number of patients	Percent female	Race ratio
14	2018	Moreira and Namen	Nortd Fluminense State University	Dementia	605	N/A	N/A
15	2012	Schipper et al	Embry-Riddle Aeronautical University	Single toxic exposures	30,152	N/A	N/A
16	2019	Giannini et al	University of Pennsylvania	Sepsis and septic shock	227,124	48.9%	52.9% White 38.4% Black 2.2% Otder
17	2017	Pestian et al	University of Cincinnati	Suicide	379	59.9%	N/A
18	2019	Thabtah et al	Manukau Institute of Technology	Autism	1,367	46.2%	35.2% White 17.0% Asian 11.8% Middle Eastern 5.6% Soutd Asian[b]
19	2002	Baxt et al	University of Pennsylvania	Myocardial infarction	2,204	60.0%	59.0% Black 24.0% White 15.7% Asian 1.3% Hispanic
20	1993	Cohen et al	New York State Institute for Basic Research in Developmental Disabilities	Autism	138	15.2%	[a]90% White
21	2019	Narayan and Satdiyamoortdy	Vellore Institute of Technology	Heart disease	6	N/A	N/A
22	2011	Sun et al	Taipei Medical University	Sleep apnea	120	N/A	N/A
23	2012	Bascil and Oztekin	Bozok University	Hepatitis	155	9.7%	N/A
24	2017	Redman et al	Baylor College of Medicine	Fatty liver disease	652	N/A	N/A
25	2015	Park and Kim	Dongguk University-Seoul	Acute appendicitis	801	53.1%	N/A
26	2018	Nemati et al	Emory University School of Medicine	Sepsis in ICU	79,527	47.2%	47.3% White 44.6% Black 1.4% Asian 0.04% Hispanic
27	2018	Shen et al	Peking University Shenzhen Graduate School	Infectious diseases	840	N/A	N/A
28	1994	Wilding et al	University of Pennsylvania	Breast and ovarian cancers	202	100%	Considerable racial mix
29	1992	Agyei-Mensah and Lin	Oklahoma State University	Sexually transmitted diseases	0	N/A	N/A
30	1994	Astion et al	University of Washington	Giant cell arteritis	807	N/A	N/A
31	2013	Seixas et al	University of Rio de Janeiro	Pleural tuberculosis	135	20.4%	N/A
32	2005	Pace et al	L. Sacco University Hospital	Gastro-esophageal reflux disease	159	55.3%	N/A
33	2018	Baldini et al	University of Pisa	Lymphoma	542	3.3%	N/A
34	2006	Hoshi et al	Tohoku Pharmaceutical University	Thyroid disease	208	67.8%	N/A

(*Continued*)

Table 2 (*Continued*)

Reference	Publication year	Author	Institution	Disease or condition	Number of patients	Percent female	Race ratio
35	2019	Murray et al	University of California	Lupus erytdematosus	1,835	[a]90%	N/A
36	2015	Hu et al	University of Minnesota	Surgical site infection	6,258	60%	83.8% White 6.6% Black 9.6% Otder
37	1992	Moneta et al	University of Genoa	Lyme borreliosis	741	N/A	N/A
38	1997	Hripcsak et al	Columbia University	Tuberculosis	450,000	N/A	N/A
39	2015	Gu et al	University of Auckland	Skin and subcutaneous tissue infections	3,886	48.8%	65.11% Pacific 11.48% Maori 5.87% European, 4.8% Middle Eastern[b]
40	2018	Karystianis et al	Macquarie University	Psychiatric evaluation	541	N/A	N/A
41	2011	Chuang	Kainan University	Liver disease	166	N/A	N/A
42	2001	Aronsky et al	Vanderbilt University	Pneumonia	742	N/A	N/A
43	2008	Polat et al	Selcuk University	Sleep apnea	83	28.9%	N/A
44	1996	Pesonen et al	University of Kuopio	Acute appendicitis	169	N/A	N/A
45	2012	Su et al	National Tsing Hua University	Pressure ulcer	168	65.4%	N/A
46	2008	Herasevich et al	Mayo Clinic	Severe sepsis and septic shock	351	51.2%	83.6% White[a] 6.9% Black 6.6% Asian 5.5% Latino or Hispanic
47	2019	Victor et al	Textsavvyapp, Inc.	Depression	671	58.0%	73.8% White 10.13% Black 8.35% Hispanic or Latino 4.47% Asian or Pacific Islander[b]
48	2016	Corey et al	Massachusetts General Hospital	Nonalcoholic fatty liver disease	1,231	55.2%	74.19% White 9.52% Black 8.71% Hispanic 1.93% Asian
49	2010	Kitporntderanunt and Wiriyasuttiwong	Srinakharinwirot University	Ectopic pregnancy	32	100%	N/A
50	2017	Mansourypoor and Asadi	University of Tehran	Diabetes	1,171	N/A	N/A
51	1998	Pesonen et al	University of Kuopio	Appendicitis	1,846	N/A	N/A
52	2000	Shang et al	University of Pittsburgh	Metdicillin-resistant *Staphylococcus aureus*	504	N/A	N/A
53	2018	Ozkan et al	Selcuk University, Konya	Urinary tract infection	59	59.3%	N/A
54	2013	Barnhart-Magen et al	Holon Institute of Technology	Thalassemia	526	N/A	100% Caucasian[a]
55	2017	Hornbrook et al		Colorectal cancer	17,095	48.8%	N/A

Table 2 (*Continued*)

Reference	Publication year	Author	Institution	Disease or condition	Number of patients	Percent female	Race ratio
			Kaiser Permanente Center for Healtd Research				
56	2016	Ng et al	IBM Research	Heart failure	15,209	49.7%	N/A
57	2018	Blecker et al	New York University School of Medicine	Acute decompensated heart failure	37,229	49.1%	10.3% Hispanic or Latino 9.9% Black
58	2017	Chase et al	Columbia University	Multiple sclerosis	2,999	72.5%	High proportion Hispanic
59	2019	Daunhawer et al	University of Basel	Neonatal hyperbilirubinemia	362	43.1%	Predominantly Caucasian
60	2019	Hu et al	Zhejiang University	Acute coronary syndrome	2,930	29%	Predominantly Asian[a]
61	1997	Viktor et al	University of Pretoria	Tuberculosis	337	N/A	N/A
62	2019	Donald et al	BrainIT Group	Arterial hypotension	104	25%	N/A
63	2015	Zhou et al	Partners Healtdcare, Inc.	Depression	1,200	N/A	N/A
64	2019	Ren et al	Shanghai Jiao Tong University	Tuberculosis pleural effusion	470	35%	100% Asian
65	2000	Vlachonikolis et al	European Institute of Healtd and Medical Sciences	Psychosis	796	61.9%	N/A
66	2017	Hao et al	Zhejiang University	Jaundice	203	N/A	N/A
67	2018	Abbas et al	Cognoa Inc.	Autism	162	20%	84% White 15% Black 1% Otder
68	2019	Matam et al	Arden University	Cardiac arrest	538	N/A	N/A
69	2019	Wilson et al	University of Michigan	Peritonsillar abscess	916	49.9%	N/A
70	2019	Masino et al	University of Pennsylvania	Early Sepsis	618	N/A	43% White 21% Black 3% Asian 3% Multiple
71	2019	Flechet et al	Katdolieke Universiteit Leuven	Acute kidney injury	252	38.4%	N/A
72	2015	Liu et al	Nanyang Technological University	Cardiac arrest	104	39.3%	67.2% Chinese 14.9% Malay 12.3% Indian 5.6% Otder
73	2019	Thirukumaran et al	University of Rochester	Surgical site infection	2,172	41%	83% White 14% Black 3% Otder
74	2018	Afzal et al	Mayo Clinic	Critical limb ischemia	792	44%	90% White
75	1997	Ellenius et al	University of Uppsala	Myocardial infarction	88	21.6%	100% Caucasian[a]
76	2010	Ibrahim et al	University of Malaya	Dengue	130	57.7%	N/A

(*Continued*)

Table 2 (*Continued*)

Reference	Publication year	Author	Institution	Disease or condition	Number of patients	Percent female	Race ratio
[77]	2011	Hsieh et al	National Yang-Ming University	Acute Appendicitis	180	53%	N/A
[78]	2016	Cook et al	Harvard Medical School	Suicide	1,453	65%	N/A
[79]	2020	Lipschuetz et al	Hadassah-Hebrew University Medical Center	Cesarean Delivery	7,473	100%	N/A
[80]	2018	Sabra et al	Oakland University	Venus tdrombosis	150	N/A	N/A
[81]	2006	Sanders et al	Vanderbilt University	Astdma	2,006	N/A	N/A
[82]	2019	Chen et al	Georgia Institute of Technology	Heart failure	34,502	49.7%	67.4% White 1.7% Black
[83]	2019	McCoy et al	Massachusetts General Hospital	Suicide	444,317	59%	75.8% White
[84]	2015	Han et al	Beijing Institute of Technology	Diabetes	7,913	N/A	N/A
[85]	2018	Teoh	Allm Inc.	Stroke	8,175	[a]50%	100% Asian[a]
[86]	1991	Baxt	University of California, San Diego Medical Center	Myocardial infarction	682	N/A	N/A
[87]	2017	Wang et al	Harvard Medical School	Stroke and major bleeding	480	N/A	N/A
[88]	2019	Corwin et al	University of Pennsylvania	Concussion	400	40%	60% Black 29% White 6% Hispanic 5% Otder
[89]	2020	Hopkins et al	Nortdwestern University Feinberg School of Medicine	Postoperative surgical site infections	4,046	52%	66% White 11% Black 11% Otder 2% Asian
[90]	2001	Wang et al	Partners HealtdCare System, Boston	Myocardial infarction	1,753	N/A	N/A
[91]	2008	Welsh et al	Mayo Clinic	Influenza	2,194	N/A	N/A

Abbreviation: N/A, not applicable.

[a] Corresponding author estimation.

[b] More than four racial groups were listed.

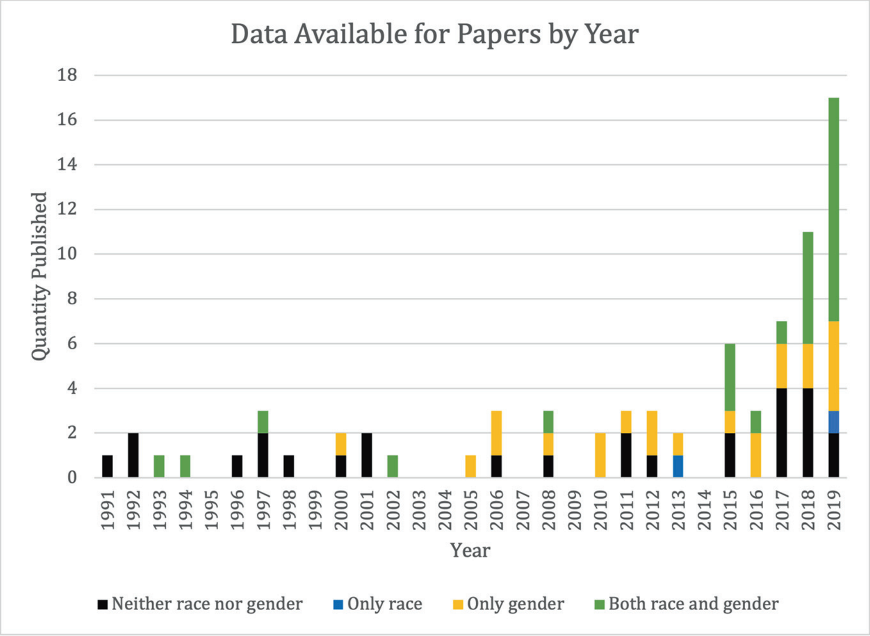

Fig. 2 Papers applying text-based machine learning to diagnosis, by data available and publish year 1991 to 2019.

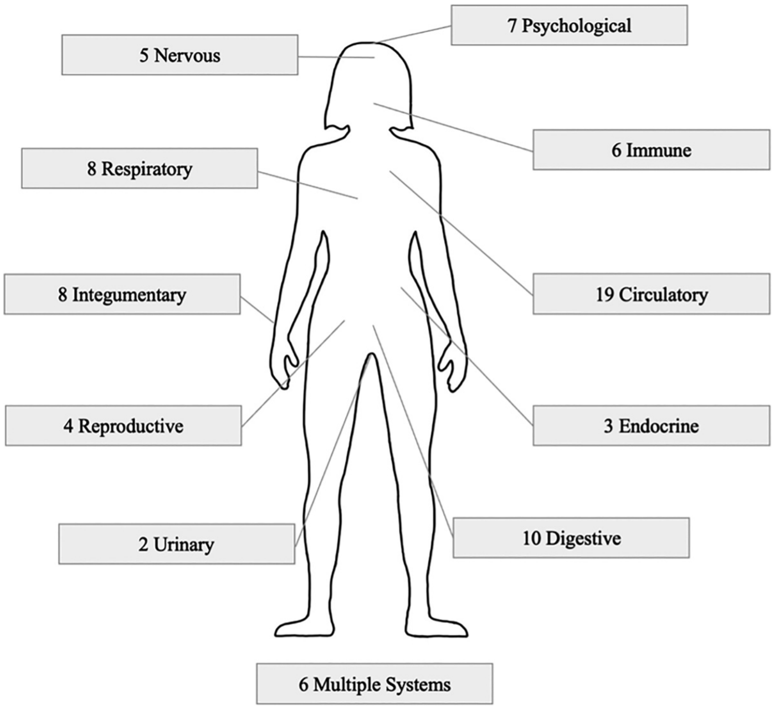

Fig. 3 Number of papers studying disease in each body system.

validation. One study included typical symptoms as determined by research and physician consult but no specific patient data.[27] When patient data were extracted from larger databases, only the patients that met the study's inclusion criteria were recorded.

Discussion

This literature review demonstrates an increasing utilization of machine learning for the analysis of text-based health information. This increase from three studies published from 1991 through 2014 to 11 studies in 2018 and 17 studies in 2019 is consistent with the shift toward reliance on informatics support in health care. As EHRs have become increasingly utilized, informatics has become more relevant in diagnostics. This is consistent with the rise in the quantity of papers published on this topic that we found. For diagnostics specifically, the availability of data in the form of EHRs is a driving force for the application of informatics.[91] Given this growing prominence, representation in the

Table 3 Diseases and conditions studied in multiple papers

Disease or condition	Number of papers
Appendicitis	4
Sepsis and septic shock	4
Autism	3
Heart failure	3
Myocardial infarction	3
Suicide	3
Cardiac arrest	2
Depression	2
Diabetes	2
Sleep apnea	2
Stroke	2
Surgical site infection	2
Tuberculosis	2

Table 4 Percentage of women in populations of included papers

Population studied	Percent of papers
Less than 50% women	59.6
Less than 30% women	20

development of predictive modeling tools is crucial to the future equity of medical diagnostics.

Training on large and diverse datasets is essential for the success of diagnostic models. With a median of 661.5 patients per study, researchers were accurately able to extract trends from large quantities of text-based data. To create robust models, however, relying on study populations with equitable demographic representation is just as relevant as incorporating clinical data from hundreds or thousands of total patients.

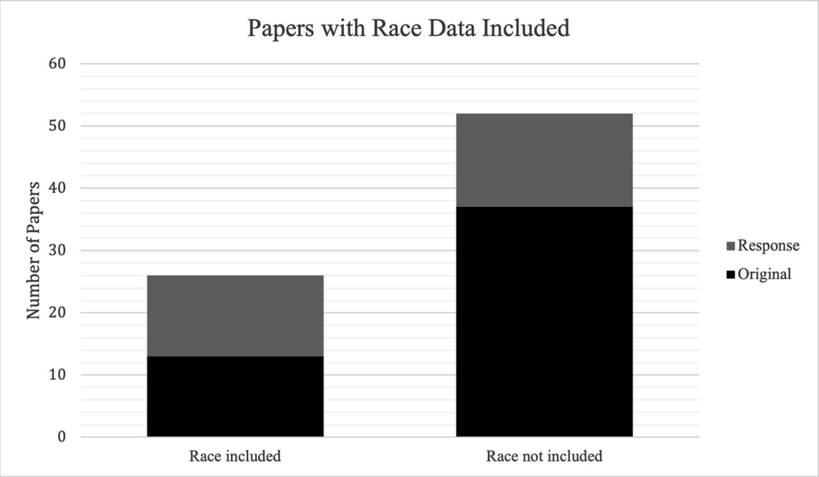

Fig. 4 Number of papers with race data included.

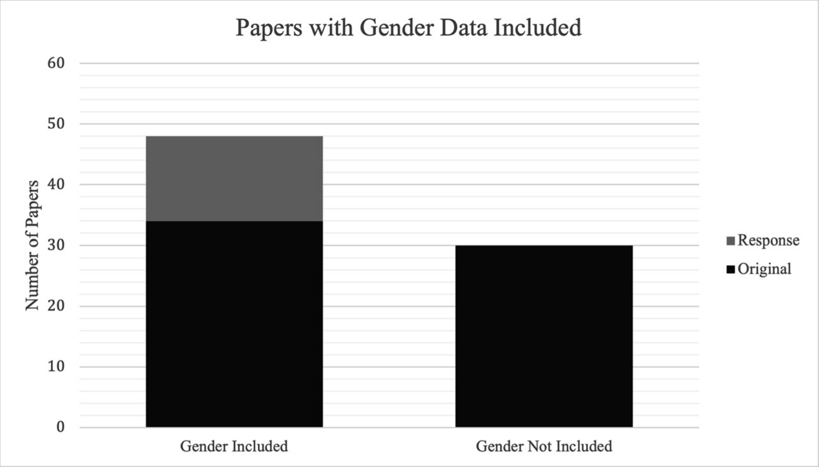

Fig. 5 Number of papers with gender data included.

The limited availability of race data was particularly alarming. While there were a variety of reasons that authors did not provide data, half of those that responded negatively specified that they never obtained these data at the time of the study. Despite recent efforts to standardize and enforce the collection of race data, this information is still chronically inaccurate or missing in the EHR.[92,93] As a result, the lack of race data provided in the papers might be largely attributed to the lack of data that were available. While documentation of race in electronic health care records is improving, it is also important for researchers to prioritize choosing data sets with study populations of which they can confirm the diversity.

To correctly report demographic information, researchers should provide deidentified data and present it in an aggregate form. Additionally, the source of both the data and the classifications used should be clearly specified. Classification categories should be as specific as possible, and it is understood that these will vary across different studies and collection formats. Category and appropriate subcategories should be listed alphabetically and reported in the results section.[94]

For the models to be generalizable to the greater populations, demographic diversity is necessary. Extrapolating to groups unrepresented in the study population leaves large gaps for potential biases. When sex, race, and ethnicity information is lacking, it is difficult to fully understand the limitations of the algorithm before expanding its use. For example, risk scores calculated on populations with limited racial and ethnic diversity have frequently been shown to perform poorly in diagnosing patients in underrepresented groups.[95–97] These issues are particularly prevalent in genome-wide association studies, as the body of previous research and genetic testing is chronically dominated by White populations.[98–100] Though race is not an ideal proxy, vulnerable populations including immigrants and those of low socioeconomic status tend to visit multiple health care facilities, resulting in health data that are more likely to be fragmented across different systems. In this way, models that rely on the quantity of encounter or the presence of an ordered test can adversely impact vulnerable populations.[101] Though there may be circumstances when training on specific populations rather than a globally representative sample is appropriate, the demographic make-up should still be well documented. For the papers that did have race data available, it was primarily a White population that was studied. Though the availability of race data is the first step, the nature of machine learning necessitates diverse study populations for diagnostic success in diverse patient populations.

An important distinction should be made between including the race of a training population in the descriptive statistics of a paper and including this feature as a component on which the machine learning algorithm relies. The belief that race accurately indicates genetic difference is antiquated, and adjusting algorithmic output based on races runs the risk of perpetuating racial biases already existing in the medical field.[97,102] This does not mean that race should be neglected altogether. Even independent of a genetic component, race, gender, and the associated social determinants of health also impact the way that patients experience disease.[103] To ensure that populations are adequately represented, these factors should be considered in the development and evaluation of machine learning algorithms.

It is important that researchers and clinicians understand how to use and access diagnostic tools.[104] The benefits of providing an accurate diagnosis are diminished if the recommendations do not have sufficient explanation. Ideally, the importance of explainable models will increase uptake and help providers make more informed decisions.[105] Models with limited interpretability, like neural networks, were relied upon most frequently. Though machine-learning-based diagnostics are becoming increasingly accurate, reliance on models that cannot be fully understood is of growing concern[95,106]

This review was limited to inclusion criteria that may not be representative of the entire breath of machine learning's integration into health care. By excluding image-based application, the scope was narrowed; however, by focusing first on diagnostics, the research can be applied to additional areas. We did not search outside of peer-reviewed literature, it is possible that studies from relevant conferences or congresses were missed. As conferences typically report incomplete work, the abstracts may not have had an impact on the results. The evidence in the review was limited by the availability of information. By contacting authors to provide additional data, other factors like how responsive a researcher was or if an email address was up to date came into play. Factors like this should be understood when considering the statistic calculated for the sex, race, and race information.

Many reviews of machine learning applications to health care do exist, yet the literature of this nature focuses largely on image-based diagnostic applications.[107–109] No literature has been found to study the availability of demographic data for papers that are both text based and diagnostic.

Conclusion

In summary, this systematic review demonstrated an increase in the application of machine learning to diagnostics in recent years. As machine learning applications gain momentum in the diagnostic field, population demographics should be carefully considered before the data can be extrapolated.

Clinical Relevance Statement

Decision support tools will continue to play an increasingly important role in clinical practice. With this, it is critical that equitable demographic representation is central to the creation and implementation of these models.

Multiple Choice Questions

1. From and EHR dataset containing records from 3,500 White men, a model is trained to successfully flag potential cases of kidney disease. What would be a primary

concern in implementing this tool into real-time clinical practice?

a. The model would be of little value as kidney disease is not difficult to diagnose.
b. The model should not be considered for implementation until it is trained on a diverse population.
c. Men would not benefit from the model as kidney disease occurs more frequently in women.
d. EHR data cannot be accessed and utilized in this way.

Correct Answer: The correct answer is option b. The model was trained exclusively on White men. The diversity of a training population is extremely significant in the generalizability of an algorithm. A model that has only been trained on White men is valuable to flag kidney disease in this demographic population, but it would be unwise to extrapolate the algorithm to different groups without first training on datasets of these populations.

2. Within the next 5 years, the reliance on artificial intelligence for clinical decision support is expected to:

a. Decrease dramatically
b. Decrease slightly
c. Remain nearly constant
d. Increase

Correct Answer: The correct answer is option d. The increase in the quantity of papers published on this topic per year indicates a trend of increasing reliance on informatics support in health care. As EHRs have become increasingly utilized, informatics has, and will continue to become, more relevant in diagnostics.

Protection of Human and Animal Subjects

Human subjects were not included in this project.

Conflict of Interest

None declared.

Acknowledgments

The authors would like to thank John Huber for providing his website to review and categorize abstracts.

References

1 EMC Digital Universe with Research and Analysis by ICD. 2014 Available at: https://www.emc.com/leadership/digital-universe/2014iview/index.htm

2 Parasrampuria S, Henry J. Hospitals' use of electronic health records data, 2015–2017. Office of the National Coordinator for Health Information Technology. 2019 Accessed April 21, 2022 at: https://www.healthit.gov/sites/default/files/page/2019-04/AHAEHRUseDataBrief.pdf

3 Murdoch TB, Detsky AS. The inevitable application of big data to health care. JAMA 2013;309(13):1351–1352

4 Kononenko I. Machine learning for medical diagnosis: history, state of the art and perspective. Artif Intell Med 2001;23(01):89–109

5 Maity NG, Das S. Machine learning for improved diagnosis and prognosis in healthcare. IEEE Aerospace Conference, 2017:1–9

6 Shah M, Shu D, Prasath VBS, Ni Y, Schapiro AH, Dufendach KR. Machine learning for detection of correct peripherally inserted central catheter tip position from radiology reports in infants. Appl Clin Inform 2021;12(04):856–863

7 Hudson DL, Cohen ME. Merging medical informatics and automated diagnostic methods. Annu Int Conf IEEE Eng Med Biol Soc 2013;2013:4783–4786

8 Zou J, Schiebinger L. AI can be sexist and racist—it's time to make it fair. Nature. Accessed October 18, 2020 at: https://www.nature.com/articles/d41586-018-05707-8?source=post_page—–817fa60d75e9

9 PubMed [database on the Internet]. Bethesda, MDNational Library of Medicine (US) Accessed April 21, 2022 at: https://pubmed.ncbi.nlm.nih.gov/

10 OVID [database on the Internet]. New York, NYOvid Technologies Accessed April 21, 2022 at: http://www.ovid.com

11 ISI Web of Knowledge [database on the Internet]. Stamford, CTThe Thompson Corporation Accessed July 13, 2020) at: http://www.isiknowledge.com

12 Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33(01):159–174

13 McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb) 2012;22(03):276–282

14 Moreira LB, Namen AA. A hybrid data mining model for diagnosis of patients with clinical suspicion of dementia. Comput Methods Programs Biomed 2018;165:139–149

15 Schipper JD, Dankel DD II, Arroyo AA, Schauben JL. A knowledge-based clinical toxicology consultant for diagnosing single exposures. Artif Intell Med 2012;55(02):87–95

16 Giannini HM, Ginestra JC, Chivers C, et al. A machine learning algorithm to predict severe sepsis and septic shock: development, implementation, and impact on clinical practice. Crit Care Med 2019;47(11):1485–1492

17 Pestian JP, Sorter M, Connolly B, et al; STM Research Group. A machine learning approach to identifying the thought markers of suicidal subjects: a prospective multicenter trial. Suicide Life Threat Behav 2017;47(01):112–121

18 Thabtah F, Abdelhamid N, Peebles D. A machine learning autism classification based on logistic regression analysis. Health Inf Sci Syst 2019;7(01):12

19 Baxt WG, Shofer FS, Sites FD, Hollander JE. A neural computational aid to the diagnosis of acute myocardial infarction. Ann Emerg Med 2002;39(04):366–373

20 Cohen IL, Sudhalter V, Landon-Jimenez D, Keogh M. A neural network approach to the classification of autism. J Autism Dev Disord 1993;23(03):443–466

21 Narayan S, Sathiyamoorthy E. A novel recommender system based on FFT with machine learning for predicting and identifying heart diseases. Neural Comput Appl 2019;31:93–102

22 Sun LM, Chiu HW, Chuang CY, Liu L. A prediction model based on an artificial intelligence system for moderate to severe obstructive sleep apnea. Sleep Breath 2011;15(03):317–323

23 Bascil MS, Oztekin H. A study on hepatitis disease diagnosis using probabilistic neural network. J Med Syst 2012;36(03):1603–1606

24 Redman JS, Natarajan Y, Hou JK, et al. Accurate identification of fatty liver disease in data warehouse utilizing natural language processing. Dig Dis Sci 2017;62(10):2713–2718

25 Park SY, Kim SM. Acute appendicitis diagnosis using artificial neural networks. Technol Health Care 2015;23(23, Suppl 2):S559–S565

26 Nemati S, Holder A, Razmi F, Stanley MD, Clifford GD, Buchman TG. An interpretable machine learning model for accurate prediction of sepsis in the ICU. Crit Care Med 2018;46(04):547–553

27 Shen Y, Yuan K, Chen D, et al. An ontology-driven clinical decision support system (IDDAP) for infectious disease diagnosis and antibiotic prescription. Artif Intell Med 2018;86:20–32

28 Wilding P, Morgan MA, Grygotis AE, Shoffner MA, Rosato EF. Application of backpropagation neural networks to diagnosis of breast and ovarian cancer. Cancer Lett 1994;77(2-3):145–153

29 Agyei-Mensah SO, Lin FC. Application of neural networks in medical diagnosis: the case of sexually-transmitted diseases. Australas Phys Eng Sci Med 1992;15(04):186–192

30 Astion ML, Wener MH, Thomas RG, Hunder GG, Bloch DA. Application of neural networks to the classification of giant cell arteritis. Arthritis Rheum 1994;37(05):760–770

31 Seixas JM, Faria J, Souza Filho JB, Vieira AF, Kritski A, Trajman A. Artificial neural network models to support the diagnosis of pleural tuberculosis in adult patients. Int J Tuberc Lung Dis 2013; 17(05):682–686

32 Pace F, Buscema M, Dominici P, et al. Artificial neural networks are able to recognize gastro-oesophageal reflux disease patients solely on the basis of clinical data. Eur J Gastroenterol Hepatol 2005;17(06):605–610

33 Baldini C, Ferro F, Luciano N, Bombardieri S, Grossi E. Artificial neural networks help to identify disease subsets and to predict lymphoma in primary Sjögren's syndrome. Clin Exp Rheumatol 201836 Suppl 112(03):137–144

34 Hoshi K, Kawakami J, Sato W, et al. Assisting the diagnosis of thyroid diseases with Bayesian-type and SOM-type neural networks making use of routine test data. Chem Pharm Bull (Tokyo) 2006;54(08):1162–1169

35 Murray SG, Avati A, Schmajuk G, Yazdany J. Automated and flexible identification of complex disease: building a model for systemic lupus erythematosus using noisy labeling. J Am Med Inform Assoc 2019;26(01):61–65

36 Hu Z, Simon GJ, Arsoniadis EG, Wang Y, Kwaan MR, Melton GB. Automated detection of postoperative surgical site infections using supervised methods with electronic health record data. Stud Health Technol Inform 2015;216:706–710

37 Moneta C, Parodi G, Rovetta S, et al. Automated diagnosis and disease characterization using neural network analysis. J Rheumatol 1995;22(03):571–572

38 Hripcsak G, Knirsch CA, Jain NL, Pablos-Mendez A. Automated tuberculosis detection. J Am Med Inform Assoc 1997;4(05): 376–381

39 Gu Y, Kennelly J, Warren J, Nathani P, Boyce T. Automatic detection of skin and subcutaneous tissue infections from primary care electronic medical records. Stud Health Technol Inform 2015;214:74–80

40 Karystianis G, Nevado AJ, Kim CH, Dehghan A, Keane JA, Nenadic G. Automatic mining of symptom severity from psychiatric evaluation notes. Int J Methods Psychiatr Res 2018;27(01):e1602

41 Chuang CL. Case-based reasoning support for liver disease diagnosis. Artif Intell Med 2011;53(01):15–23

42 Aronsky D, Fiszman M, Chapman WW, Haug PJ. Combining decision support methodologies to diagnose pneumonia. Proc AMIA Symp 2001:12–16

43 Polat K, Yosunkaya S, Güneş S Comparison of different classifier algorithms on the automated detection of obstructive sleep apnea syndrome. J Med Syst 2008;32(03):243–250

44 Pesonen E, Eskelinen M, Juhola M. Comparison of different neural network algorithms in the diagnosis of acute appendicitis. Int J Biomed Comput 1996;40(03):227–233

45 Su CT, Wang PC, Chen YC, Chen LF. Data mining techniques for assisting the diagnosis of pressure ulcer development in surgical patients. J Med Syst 2012;36(04):2387–2399

46 Herasevich V, Afessa B, Chute CG, Gajic O. Designing and testing computer based screening engine for severe sepsis/septic shock. AMIA Annu Symp Proc 2008;966:966

47 Victor E, Aghajan ZM, Sewart AR, Christian R. Detecting depression using a framework combining deep multimodal neural networks with a purpose-built automated evaluation. Psychol Assess 2019;31(08):1019–1027

48 Corey KE, Kartoun U, Zheng H, Shaw SY. Development and validation of an algorithm to identify nonalcoholic fatty liver disease in the electronic medical record. Dig Dis Sci 2016;61(03): 913–919

49 Kitporntheranunt M, Wiriyasuttiwong W. Development of a medical expert system for the diagnosis of ectopic pregnancy. J Med Assoc Thai 2010;93(Suppl 2):S43–S49

50 Mansourypoor F, Asadi S. Development of a reinforcement learning-based evolutionary fuzzy rule-based system for diabetes diagnosis. Comput Biol Med 2017;91:337–352

51 Pesonen E, Ohmann C, Eskelinen M, Juhola M. Diagnosis of acute appendicitis in two databases. Evaluation of different neighborhoods with an LVQ neural network. Methods Inf Med 1998;37 (01):59–63

52 Shang JS, Lin YS, Goetz AM. Diagnosis of MRSA with neural networks and logistic regression approach. Health Care Manage Sci 2000;3(04):287–297

53 Ozkan IA, Koklu M, Sert IU. Diagnosis of urinary tract infection based on artificial intelligence methods. Comput Methods Programs Biomed 2018;166:51–59

54 Barnhart-Magen G, Gotlib V, Marilus R, Einav Y. Differential diagnostics of thalassemia minor by artificial neural networks model. J Clin Lab Anal 2013;27(06):481–486

55 Hornbrook MC, Goshen R, Choman E, et al. Early colorectal cancer detected by machine learning model using gender, age, and complete blood count data. Dig Dis Sci 2017;62(10): 2719–2727

56 Ng K, Steinhubl SR, deFilippi C, Dey S, Stewart WF. Early detection of heart failure using electronic health records: practical implications for time before diagnosis, data diversity, data quantity, and data density. Circ Cardiovasc Qual Outcomes 2016;9(06): 649–658

57 Blecker S, Sontag D, Horwitz LI, et al. Early identification of patients with acute decompensated heart failure. J Card Fail 2018;24(06):357–362

58 Chase HS, Mitrani LR, Lu GG, Fulgieri DJ. Early recognition of multiple sclerosis using natural language processing of the electronic health record. BMC Med Inform Decis Mak 2017;17(01):24

59 Daunhawer I, Kasser S, Koch G, et al. Enhanced early prediction of clinically relevant neonatal hyperbilirubinemia with machine learning. Pediatr Res 2019;86(01):122–127

60 Hu D, Dong W, Lu X, Duan H, He K, Huang Z. Evidential MACE prediction of acute coronary syndrome using electronic health records. BMC Med Inform Decis Mak 2019;19(Suppl 2):61

61 Viktor HL, Cloete I, Beyers N. Extraction of rules for tuberculosis diagnosis using an artificial neural network. Methods Inf Med 1997;36(02):160–162

62 Donald R, Howells T, Piper I, et al; BrainIT Group. Forewarning of hypotensive events using a Bayesian artificial neural network in neurocritical care. J Clin Monit Comput 2019;33(01):39–51

63 Zhou L, Baughman AW, Lei VJ, et al. Identifying patients with depression using free-text clinical documents. Stud Health Technol Inform 2015;216:629–633

64 Ren Z, Hu Y, Xu L. Identifying tuberculous pleural effusion using artificial intelligence machine learning algorithms. Respir Res 2019;20(01):220

65 Vlachonikolis IG, Karras DA, Hatzakis MJ, Paritsis N. Improved statistical classification methods in computerized psychiatric diagnosis. Med Decis Making 2000;20(01):95–103

66 Hao SR, Geng SC, Fan LX, Chen JJ, Zhang Q, Li LJ. Intelligent diagnosis of jaundice with dynamic uncertain causality graph model. J Zhejiang Univ Sci B 2017;18(05):393–401

67 Abbas H, Garberson F, Glover E, Wall DP. Machine learning approach for early detection of autism by combining questionnaire and home video screening. J Am Med Inform Assoc 2018;25(08):1000–1007

68 Matam BR, Duncan H, Lowe D. Machine learning based framework to predict cardiac arrests in a paediatric intensive care unit: prediction of cardiac arrests. J Clin Monit Comput 2019;33(04): 713–724

69 Wilson MB, Ali SA, Kovatch KJ, Smith JD, Hoff PT. Machine learning diagnosis of peritonsillar abscess. Otolaryngol Head Neck Surg 2019;161(05):796–799

70 Masino AJ, Harris MC, Forsyth D, et al. Machine learning models for early sepsis recognition in the neonatal intensive care unit using readily available electronic health record data. PLoS One 2019;14(02):e0212665

71 Flechet M, Falini S, Bonetti C, et al. Machine learning versus physicians' prediction of acute kidney injury in critically ill adults: a prospective evaluation of the AKIpredictor. Crit Care 2019;23(01):282

72 Liu T, Lin Z, Ong ME, et al. Manifold ranking based scoring system with its application to cardiac arrest prediction: a retrospective study in emergency department patients. Comput Biol Med 2015;67:74–82

73 Thirukumaran CP, Zaman A, Rubery PT, et al. Natural language processing for the identification of surgical site infections in orthopaedics. J Bone Joint Surg Am 2019;101(24):2167–2174

74 Afzal N, Mallipeddi VP, Sohn S, et al. Natural language processing of clinical notes for identification of critical limb ischemia. Int J Med Inform 2018;111:83–89

75 Ellenius J, Groth T, Lindahl B. Neural network analysis of biochemical markers for early assessment of acute myocardial infarction. Stud Health Technol Inform 1997;43(Pt A):382–385

76 Ibrahim F, Faisal T, Salim MI, Taib MN. Non-invasive diagnosis of risk in dengue patients using bioelectrical impedance analysis and artificial neural network. Med Biol Eng Comput 2010;48 (11):1141–1148

77 Hsieh CH, Lu RH, Lee NH, Chiu WT, Hsu MH, Li YC. Novel solutions for an old disease: diagnosis of acute appendicitis with random forest, support vector machines, and artificial neural networks. Surgery 2011;149(01):87–93

78 Cook BL, Progovac AM, Chen P, Mullin B, Hou S, Baca-Garcia E. Novel use of natural language processing (NLP) to predict suicidal ideation and psychiatric symptoms in a text-based mental health intervention in Madrid. Comput Math Methods Med 2016;2016:8708434

79 Lipschuetz M, Guedalia J, Rottenstreich A, et al. Prediction of vaginal birth after cesarean deliveries using machine learning. Am J Obstet Gynecol 2020;222(06):613.e1–613.e12

80 Sabra S, Mahmood Malik K, Alobaidi M. Prediction of venous thromboembolism using semantic and sentiment analyses of clinical narratives. Comput Biol Med 2018;94:1–10

81 Sanders DL, Aronsky D. Prospective evaluation of a Bayesian network for detecting asthma exacerbations in a pediatric emergency department. AMIA Annu Symp Proc 2006;2006:1085

82 Chen R, Stewart WF, Sun J, Ng K, Yan X. Recurrent neural networks for early detection of heart failure from longitudinal electronic health record data: implications for temporal modeling with respect to time before diagnosis, data density, data quantity, and data type. Circ Cardiovasc Qual Outcomes 2019;12 (10):e005114

83 McCoy TH Jr, Pellegrini AM, Perlis RH. Research domain criteria scores estimated through natural language processing are associated with risk for suicide and accidental death. Depress Anxiety 2019;36(05):392–399

84 Han L, Luo S, Yu J, Pan L, Chen S. Rule extraction from support vector machines using ensemble learning approach: an application for diagnosis of diabetes. IEEE J Biomed Health Inform 2015; 19(02):728–734

85 Teoh D. Towards stroke prediction using electronic health records. BMC Med Inform Decis Mak 2018;18(01):127

86 Baxt WG. Use of an artificial neural network for the diagnosis of myocardial infarction. Ann Intern Med 1991;115(11):843–848

87 Wang SV, Rogers JR, Jin Y, Bates DW, Fischer MA. Use of electronic healthcare records to identify complex patients with atrial fibrillation for targeted intervention. J Am Med Inform Assoc 2017;24(02):339–344

88 Corwin DJ, Propert KJ, Zorc JJ, Zonfrillo MR, Wiebe DJ. Use of the vestibular and oculomotor examination for concussion in a pediatric emergency department. Am J Emerg Med 2019;37 (07):1219–1223

89 Hopkins BS, Mazmudar A, Driscoll C, et al. Using artificial intelligence (AI) to predict postoperative surgical site infection: a retrospective cohort of 4046 posterior spinal fusions. Clin Neurol Neurosurg 2020;192:105718

90 Wang SJ, Ohno-Machado L, Fraser HS, Kennedy RL. Using patient-reportable clinical history factors to predict myocardial infarction. Comput Biol Med 2001;31(01):1–13

91 Welsh G, Wahner-Roedler D, Froehling DA, Trusko B, Elkin P. Whole record surveillance is superior to chief complaint surveillance for predicting influenza. AMIA Annu Symp Proc 2008; 1173:1173

92 Polubriaginof FCG, Ryan P, Salmasian H, et al. Challenges with quality of race and ethnicity data in observational databases. J Am Med Inform Assoc 2019;26(8-9):730–736

93 Sholle ET, Pinheiro LC, Adekkanattu P, et al. Underserved populations with missing race ethnicity data differ significantly from those with structured race/ethnicity documentation. J Am Med Inform Assoc 2019;26(8-9):722–729

94 Flanagin A, Frey T, Christiansen SLAMA Manual of Style Committee. Updated guidance on the reporting of race and ethnicity in medical and science journals. JAMA 2021;326(07):621–627

95 Parikh RB, Teeple S, Navathe AS. Addressing bias in artificial intelligence in health care. JAMA 2019;322(24):2377–2378

96 Gijsberts CM, Groenewegen KA, Hoefer IE, et al. Race/ethnic differences in the associations of the Framingham risk factors with carotid IMT and cardiovascular events. PLoS One 2015;10 (07):e0132321

97 Powe NR. Black kidney function matters: use or misuse of race? JAMA 2020;324(08):737–738

98 Weinberger DR, Dzirasa K, Crumpton-Young LL. Missing in action: African ancestry brain research. Neuron 2020;107(03):407–411

99 McCarthy AM, Bristol M, Domchek SM, et al. Health care segregation, physician recommendation, and racial disparities in BRCA1/2 testing among women with breast cancer. J Clin Oncol 2016;34(22):2610–2618

100 Suther S, Kiros GE. Barriers to the use of genetic testing: a study of racial and ethnic disparities. Genet Med 2009;11(09): 655–662

101 Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G. Potential biases in machine learning algorithms using electronic health record data. JAMA Intern Med 2018;178(11):1544–1547

102 Vyas DA, Eisenstein LG, Jones DS. Hidden in plain sight—reconsidering the use of race correction in clinical algorithms. N Engl J Med 2020;383(09):874–882

103 Bamshad M. Genetic influences on health: does race matter? JAMA 2005;294(08):937–946 [Erratum in: JAMA. 2005 Oct 5;294(13):1620. PMID: 16118384]

104 Liu X, Anstey J, Li R, Sarabu C, Sono R, Butte AJ. Rethinking PICO in the machine learning era: ML-PICO. Appl Clin Inform 2021;12 (02):407–416

105 Adlung L, Cohen Y, Mor U, et al. Machine learning in clinical decision making. Med 2021;2(06):642–665

106 Holzinger A, Langs G, Denk H, Zatloukal K, Müller H. Causability and explainability of artificial intelligence in medicine. Wiley Interdiscip Rev Data Min Knowl Discov 2019;9(04):e1312

107 Thomsen K, Iversen L, Titlestad TL, Winther O. Systematic review of machine learning for diagnosis and prognosis in dermatology. J Dermatolog Treat 2020;31(05):496–510

108 de Filippis R, Carbone EA, Gaetano R, et al. Machine learning techniques in a structural and functional MRI diagnostic approach in schizophrenia: a systematic review. Neuropsychiatr Dis Treat 2019;15:1605–1627

109 Kassem MA, Hosny KM, Damaševičius R, Eltoukhy MM. Machine learning and deep learning methods for skin lesion classification and diagnosis: a systematic review. Diagnostics (Basel) 2021;11 (08):1390