BMJ Health & Care Informatics

Women's health service access and associated factors in Ethiopia: application of geographical information system and multilevel analysis

Addisalem Workie Demsash , Agmasie Damtew Walle

To cite: Demsash AW, Walle AD. Women's health service access and associated factors in Ethiopia: application of geographical information system and multilevel analysis. *BMJ Health Care Inform* 2023;**30**:e100720. doi:10.1136/bmjhci-2022-100720

Received 17 December 2022
Accepted 12 April 2023

Mettu University, College of Health Science, Health Informatics Department, Mettu, Ethiopia

Correspondence to
Addisalem Workie Demsash;
addisalemworkie599@gmail.com

ABSTRACT

Objectives Women's access to healthcare services is challenged by various factors. This study aimed to assess women's health service access and identify associated factors.

Methods A cross-sectional study design with a two-stage stratified sampling technique, and 12 945 women from the 2016 Ethiopia Demographic and Health Survey dataset were used. The spatial hotspot analysis and purely Bernoulli-based model scan statistics were used to highlight hot and cold spot areas, and to detect significant local clusters of women's health service access. A multilevel logistic regression analysis was used to assess factors that affect women's access to health services. A variable with a p<o.o5 was considered as a significant factor.

Results Overall, 29.8%% of women had health services access. 70.2% of women had problems with health services access such as: not wanting to go alone (42%), distance to health facilities (51%), getting the money needed for treatment (55%) and getting permission to go for medical care (32.3%). The spatial distribution of health service access in Ethiopia was clustered, and low health service access was observed in most areas of the country. Women who lived in primary, secondary and tertiary clusters were 96%, 39% and 72% more likely to access health services. Educational status, rich wealth status, media exposure and rural residence were statistically significant factors.

Conclusions In Ethiopia, women have problems with health services access. The spatial distribution of health services access was non-random, and hotspot areas of women's health service access were visualised in parts of Benishangul Gumez, Amhara, Afar, DireDawa, Harari, and Somali regions. Creating job opportunities, public health promotion regarding maternal health service utilisation and constructing nearby health facilities are required for better healthcare service access for women.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Access to maternal healthcare services remains a significant problem in the world. Women's health service access is poor in low-income and middle-income countries and maternal and child deaths are high.

WHAT THIS STUDY ADDS

⇒ This study used nationally representative data. Application of geographical information system was applied, and women's health service access was spatially presented and visualised through maps. Factors that affect women's health service access were identified.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ The findings are crucial for stakeholders to give priority attention to the cold spot areas of women's health service access, and create awareness for women regarding health service access. This study illustrates the practical application of geographical information systems to concerned women's health service access, and the findings are used as a basis for future studies.

INTRODUCTION

Globally, around 800 women died daily from preventable causes, pregnancy and childbirth.[1] Worldwide, nearly 300 000 women were predicted to be died in 2010.[2] Maternal and child mortality remains a major public health challenge in the developing world, and the discrepancy is high according to developed countries. The maternal mortality ratio in low-income and middle-income countries is 15 times higher than in developed countries.[3] Sub-Saharan Africa had the highest maternal mortality ratio which is 64 deaths per 10 000 live births, and 900 maternal deaths occurred in Ethiopia.[3] According to the 2011 Ethiopian Demographic and Health Survey (EDHS), maternal mortality rates were 6.76 per 1000 live births.[4]

Maternal and child mortality occur due to countries' poor utilisation of healthcare services.[4] An average of 52% of women received at least four antenatal care (ANC) services in developing regions. It was 36% in Asia, 49% in sub-Saharan Africa[5] and 32% in Ethiopia.[6] In addition, 77.69%, 73.95% and 67.61% of women delayed their first ANC visit in 2005, 2011 and 2016 EDHS, respectively.[7] Pregnancy-induced high blood pressure, stillbirth and unsafe abortions are extremely high. Though institutional delivery is used to reduce pregnancy and birth risks, home birth in low-income countries is high and institutional delivery is 26%–32.5% in Ethiopia.[8]

Women's postnatal care (PNC) service utilisation is insufficient, which is 6.9% with marked spatial heterogeneity in Ethiopia.[9] Maternal and child health problems occur due to the inaccessibility of health institutions, poor women's health-seeking behaviours, maternal and child health services inaccessibility, low media exposure, poor attitude and knowledge, and low-quality service provision.[8 10 11] The geographical accessibility of healthcare services, rural women lack reliable transportation, and women are late in starting ANC service.[12]

In resource-limited settings, women's access to health services is mainly affected by four principal factors such as not want to go alone, distance to health facilities, getting the money needed for treatment, and getting permission for medical care. Previous studies have proven that distance to a health facility (long distance) and the geographical position of health facilities,[13] poverty, low monthly income and not having an occupation that makes women not have enough money for medical care,[12] inadequate awareness and low-risk perception of women, laziness and disease severity that make women not want to go alone for medical care,[14] and husbands' and relatives' complete decision-making culture for women's health service access[15 16] are challenges for maternal and child health service access among women.

Maternal and child healthcare services are the most effective and potential health interventions to overcome maternal and child mortality. Of these maternal and child healthcare services, the provision of ANC and PNC services,[17] institutional delivery services,[18] skilled birth assistance, and nutritional and breastfeeding counselling services are the main strategies of the Millennium Development Goal to reduce mortality and morbidity. Moreover, health facilities provide preventative, curative health services to prevent maternal and child deaths.[9]

Maternal health services would be accessible and fairly distributed,[19] the quality of health service provision should be ensured[20] and sufficient health professionals would available in health facilities. Plus, policy-makers understand women's health services access problem and geographical variations of health service access to formulate strategies and interventions to solve women's health services problems and to provide equity and quality of services provision.[17 21] Therefore, this study would be an input for policy-makers to alleviate women's health service access problems. Studies regarding women's health service access are not adequate, and the findings of previous studies were insufficient and limited in spatial variation analysis of women's health service access in Ethiopia. Moreover, the findings of this study would be important for women's decision-making regarding maternal health service access, and the finding could support policy-makers, and programmers to design interventions for achieving women's health service access. Therefore, this study aimed to assess women's health services access, locate women's health service access spatially and identify factors associated with women's health service access.

METHODS

Study design

A cross-sectional study design was used.

Study setting

The study was conducted across nine regions of Ethiopia. Ethiopia has nine regional states with two city administrations. The country is located in the Horn of Africa and is bordered by Eritrea to the north, Djibouti, and Somalia to the east, Sudan and South Sudan to the west, and Kenya to the South.

Data sources

The 2016 EDHS dataset was used. The 2016 EDHS was the fourth Demographic and Health Survey (DHS) conducted in Ethiopia. The survey was conducted by Ethiopian Public Health Institute at the request of the Federal Ministry of Health. According to the Ethiopian EDHS report, the survey was conducted from 18 January 2016 to 27 June 2016. The actual data for this study were accessed from the measure DHS website (www.dhsprogram.com), and Ethiopian shape files were downloaded from the open Africa website (https://africaopendata.org/dataset).

Sampling producers

The 2016 EDHS was conducted in 2007 by the Central Statistical Agency. The census frame is the complete list of 84915 enumeration areas that cover an average of 181 households. A two-stage stratified cluster sampling was used, each region was stratified into urban and rural areas. At the first stage of selection, a total of 645 enumeration areas were selected independently with probability proportion to each enumeration areas. In second stage of selection, a fixed number of 28 households per cluster were selected with an equal probability of systematic selection from the newly created household listing. For more detail about the methods, visit the 2016 EDHS report.[22]

Study populations

All women who were either permanent residents of the selected households or visitors who stayed in the household the night before the survey were the source population. Whereas, all women aged 15–49 years were the study population. Zero coordinates areas, clusters that had no defined proportions of health service access and irregularly shaped clusters were excluded.

Variables of the study

Dependent variable

Women's health service access.

Women

In this study, women are all eligible women aged 15–49 years who are a permanent resident or visitors of the selected households available before the survey interview begin.

Women's health services access was challenged by different factors. In this study, different factors that challenge women's health services access were adapted from the EDHS report and other similar studies.[13] Accordingly, women's health service access was assessed regardless of the following four factors such as (1) not wanting to go alone, (2) distance to health facilities, (3) getting the money needed for treatment and (4) getting permission to go for medical care. Hence, the women had health service access if they had not been faced with any of the mentioned factors. Otherwise, the women had problems with health service access if they faced by at least one of the mentioned factors.

Independent variables

Age, educational status, wealth status, religion, media exposure and current working were used as individual-label independent variables. Region and place of residency were used as community-label variables.

Media exposure

Women's health service access is related to their media exposure. Therefore, if the women had either radio, television or both the women had media exposure, and if the women had not had either radio or television, the women had no media exposure.[23]

Data management and processing

Data cleaning, labelling and processing were done by using STATA V.15 software and Microsoft Office Excel. To yield accurate parameters estimation, and to handle the representativeness of the survey, sample weight was done.

Statically data analysis

STATA V.15 software was used for data processing and analysis. A descriptive analysis was done to describe the characteristics of the study subjects.

Spatial data analysis

ArcMap V.10.7 software was used for spatial autocorrelation and hot spot analysis. Global spatial autocorrelation (Global Moran's I) statistic measure was used to assess whether women's health service access was dispersed, clustered or randomly distributed. Moran's I value close to −1, +1 and 0 indicates a dispersed, clustered and random distribution of health service access, respectively.[24] Hot spot analysis (Getis-Ord Gi*) was done to know whether the women's health service access is a hot or cold spot. The hot and cold spot values for spatial clusters were determined by z-scores and p values.

Spatial interpolation

We used the ordinary Kriging interpolation technique to predict health service access in the unsampled EAs. Health service access in the unsampled EAs was predicted by interpolating the currently sampled areas.

Spatial scan statistics

SaTScan V.9.5 software was used for the local cluster detection. Purely spatial Bernoulli-based models were employed to determine statistically significant clusters with high rates of women's health service access.[25] The women who were not faced health service access problems were taken as cases and those who had health service access problems were taken as controls to fit the purely spatial Bernoulli model. The default maximum spatial cluster size of less than 50% of the population was used as an upper limit, and to allow small and large clusters to be detected, and to ignore the clusters that exist the outside the maximum limits of the circular shape of the window. A log-likelihood ratio (LLR) test statistic was used to determine whether the number of observed cases within a cluster was significantly higher than expected or not. The circle with a maximum LLR was defined as the most likely (a primary) cluster. Then, all the remaining significant clusters were ranked based on their LLR.[26] All most likely significant clusters were identified using p values, and ranked by their LLR test based on the 9999 Monte Carlo replications.[26]

Multilevel mixed effect logistic regression analysis

Since the EDHS data had a hierarchical nature. Hence, women from the same cluster are more similar as compared with women who were from different clusters. Such kind of hierarchy of data might have a dependency nature. Therefore, this may violate the independence of observations and the equal variance of assumption. To overcome this violation, multilevel mixed-effect logistic regression models were assumed, and four models were considered to overcome if there is any data dependency: model 1 (a null model), model 2 (contains individual-level variables), model 3 (contains community-level variables) and model 4 (individual and community-level variables). For each model, the intraclass correlation coefficient (ICC) and variance were calculated to check the presence of data dependency and to apply multilevel mixed-effect logistic regression. ICC is used to the diagnosed correlation between clusters, and there are data correlation if ICC's value is greater than 25%. Consequently, 40% of the ICC's values confirmed that there was a significant correlation among women regarding their response to health service access. The LLR was used for model comparison, and the model with the highest LLR value was chosen as the best-fit model.[27] As a result, model D was chosen as the best-fit model due to its LLR score's highest value (−2598.4) as compared with other models (table 3). In addition, therefore, a multilevel mixed effect logistic regression analysis was fitted. A $p<0.05$ and a 95% CI were used to identify associated factors.

RESULTS

Sociodemographic characteristics of the study participants

A total of 15 295 women (weighted) were included. More than one-third (36.7%) of women were from the Oromia region. The majority (77.8%) of women were rural residents. Nearest to half (48.2%) of women had no formal education. Two out of 10 women (21.3%) were under 15–19 years of age. Forty-six per cent of women were rich. Four out of 10 women (42.8%) were orthodox religious flowers. The majority (66.6%) of the women were not employed (table 1).

Table 1 Sociodemographic characteristics of the study participants

Variable	Category	Frequency (n)	%
Educational status of mother/ caregiver	No formal education	7379	48.2
	Primary	5367	35.1
	Secondary	1721	11.3
	Higher	828	5.4
Region	Tigray	1099	7.2
	Afar	126	0.8
	Amhara	3533	23.1
	Oromia	5613	36.7
	Somali	457	3.0
	Benishangul	158	1.0
	SNNPR	3245	21.2
	Gambela	43	0.3
	Harari	38	0.2
	Addis Adaba	896	5.9
	Dire Dawa	88	0.6
Respondents' age (year)	15–19	3259	21.3
	20–24	2655	17.4
	25–29	2893	18.9
	30–34	2299	15.0
	35–39	1911	12.5
	40–44	1278	8.4
	45–49	1002	6.6
Family's wealth index	Poor	5339	34.9
	Middle	2914	19.0
	Rich	7043	46.0
Mother/caregiver religion	Orthodox	6545	42.8
	Catholic	120	0.8
	Protestant	3624	23.7
	Muslin	4797	31.4
	Traditional	123	0.8
Place of residency	Urban	3389	22.2
	Rural	11 906	77.8
Currently working	No	10 187	66.6
	Yes	5108	33.4

SNNPR, South Nations Nationalities and People's Region.

Spatial distribution of women's health service access

The women were assessed whether they had problems regarding health service access or not. Accordingly, women had a problem with not wanting to go alone (42%), distance to health facilities (51%), getting the money needed for treatment (55%) and getting permission to go for medical care (32.3%). Overall, 70.2% of women had at least one of the mentioned problems for health service access in Ethiopia (figure 1).

The spatial autocorrelation revealed that the spatial distribution of health service access in Ethiopia was clustered (Global Moran's I=0.102168, p=0.034569). These hot spots of health service access were observed in eastern Benishangul Gumuz, western Amhara, southern Afar, DireDawa, Harari and northern Somali regions (figures 2 and 3).

Spatial SaTScan analysis

A total of 72 significant clusters were identified. Of these, 9, 62 and 1 cluster were primary, secondary and tertiary clusters, respectively. The primary and secondary clusters were located at 9.614701N, 41.829121E within a 5.69 km radius in Dire Dawa, and at 10.333829 N, 34.842459 E within a 386.61 km radius in Gambela, Benishangul-Gumuz, western Oromia and southwestern Amhara regions, respectively. Women who lived in the primary, and secondary clusters were 96% (RR=1.96, p<0.0001), and 39% (RR=1.39, p<0.0001) more likely to access health service than women who lived outside the window (table 2, figure 4).

Interpolation of women's health service access

The kriging interpolation of women's health service access revealed that there would be good health service access among women in Benishangul-Gumuz, western Amhara, Dire Dawa, eastern Oromia and northern Somali regions of Ethiopia. Whereas, women in the remaining parts of Ethiopia would face problems with health service access (figure 5).

Factors associated with women's health service access

In multilevel mixed-effect logistic regression analysis; education, wealth status, media exposure and residency were significant factors for women's health service access.

The women who were in secondary, and higher education were 1.6 (adjusted odds ratio (AOR): 1.56, 95% CI 1.34 to 1.81), and 2 (AOR 2.02, 95% CI 1.66 to 2.44) times more likely to access health services than women who had no formal education. Rich women were 1.4 (AOR 1.38, 95% CI 1.19 to 1.61) times more likely to access health services than poor women. The women who had media exposure were 1.2 (AOR 1.15, 95% CI 1.03 to 1.29) times more likely to access health services than their counterparts. Rural women were 82% (AOR 0.18, 95% CI 0.14

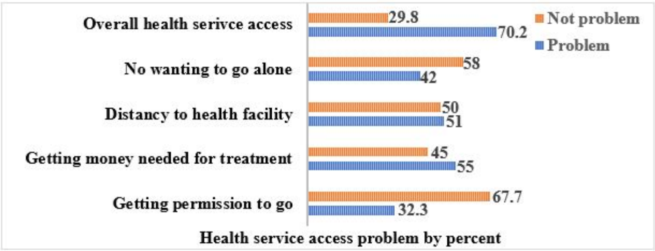
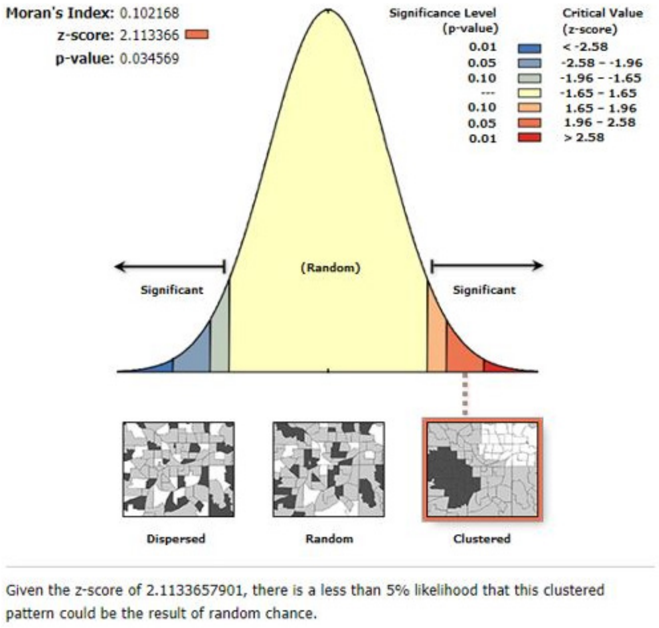


Figure 1 Health service access problem in Ethiopia, 2016 EDHS. EDHS, Ethiopian Demographic and Health Survey.

to 0.23) less likely to access health services than urban resident women (table 3).

DISCUSSION

For this study, 2016 EDHS data were used that was accessed from the DHS website. Through request, permission was obtained to access the data. There are no attributes that uniquely identify individuals' women or household addresses in the data files. This is because the geographical coordinate files are randomly displaced within a large geographical area, and it is only for EAs as a whole. As a result, specific enumeration areas (ERs), individuals' women and households cannot be identified uniquely. The shape file of Ethiopia was taken from the open Africa website. A two-stage stratified cluster sampling technique was used, and all women under the age of 15–49 years were the study population. Since the data have hierarchical nature, data dependency might have existed. Therefore, ICC was used to assess data dependency. Based on the result, multilevel mixed-effect logistic regression models were considered to alleviate the data dependency, and different model selection criteria were assumed to select the best-fit model. For spatial analysis, spatial autocorrelation and hot spot analysis were used to assess the distribution of data, and identify the hot or cold spot areas of women's health service access, respectively. The ordinary Kriging interpolation technique and purely spatial Bernoulli model were used to predict unsampled areas, and to detect local clusters of women's health service access, respectively.



Given the z-score of 2.1133657901, there is a less than 5% likelihood that this clustered pattern could be the result of random chance.

Figure 2 Spatial autocorrelation report of health service access in Ethiopia, 2016 EDHS. EDHS=Ethiopian Demographic and Health Survey.

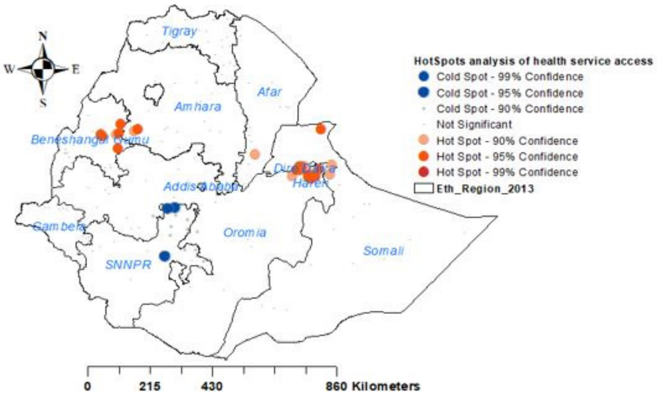

N
W
E
S
Tigray
Amhara
Afar
Beneshangul Gumu
Dire Dawa
Harer
Addis Ababa
Gambela
Oromia
SNNPR
Somali
HotSpots analysis of health service access
Cold Spot - 99% Confidence
Cold Spot - 95% Confidence
Cold Spot - 90% Confidence
Not Significant
Hot Spot - 90% Confidence
Hot Spot - 95% Confidence
Hot Spot - 99% Confidence
Eth_Region_2013
0
215
430
860 Kilometers

Figure 3 Hot spot analysis for health service access in Ethiopia, 2016 EDHS. EDHS, Ethiopian Demographic and Health Survey; SNNPR, South Nations Nationalities and People's Region.

Women's health service access was assessed to determine whether they had problems regarding health service access or not. Accordingly, respondents had problems getting the money needed for treatment (55%), distance to health facilities (51%), not wanting to go alone (42%) and getting permission to go for medical care (32.3%). Overall, 70.2% of women had at least 1 of the mentioned problems with health services access, and only 18.9% of women had good health service access in Ethiopia. This evidence was supported by studies done in Nigeria[11] and Ethiopia.[28] This finding was also supported by women's suboptimal ANC visits in Ethiopia, which ranged from 10.0% to 32%,[29] and low utilisation of PNC service utilisation.[9] This might be because women living far from health facilities are less likely to use or access healthcare services, their poor perception of the available healthcare services and lack of transportation services. In addition, mothers might not know about signs of pregnancy complications, women's low health-seeking behaviours, inaccessibility of health institutions and women's low ANC and PNC visits.[11] Therefore, stakeholders create awareness for women to make them volunteer to go alone for medical care, and husbands and other relatives might prevent women to go to the health facility and access the respective health service. So, awareness is also created for husbands and relatives not to prevent women to access health services. Furthermore, nearby health facility for women is critical to ensure equal health service access and to meet the target of maternal and child healthcare services utilisation. As well as policy-makers should enhance the economic status of women.

Table 2 SaTScan analysis report of significant clusters for women's health service access in the detected window in Ethiopia, using 2016 EDHS data

Types of cluster	Detected cluster	Coordinates/radius	Populations	Case	RR	LLR	P value
Primary	282, 285, 287, 286, 297, 296, 292, 290, 293	(9.614701N,41.829121E) 5.69 km	318	198	1.96	58.59	<0.001
Secondary	149, 151, 153, 152, 147, 154, 155, 160, 158, 170, 161, 159, 169, 167, 148, 168, 164, 163, 118, 162, 92, 80, 120, 79, 218, 211, 208, 94, 230, 229, 217, 220, 98, 53, 213, 52, 214, 206, 54, 81, 76, 97, 70, 59, 225, 85, 226, 221, 223, 210, 57, 71, 84, 96, 73, 99, 91, 95, 195, 201, 200, 112	(10.333829N,34.842459E)/ 386.61 km	1732	723	1.39	37.67	<0.001
Tertiary	247	(9.287253N, 42.135531 E)/0 km	26	21	2.43	12.39	<0.001

EDHS, Ethiopian Demographic and Health Survey; LLR, log-likelihood ratio; RR, relative risk.

The spatial distribution of health service access in Ethiopia was not random. High health service access was observed in eastern Benishangul Gumuz, southwest Amhara, southern Afar, DireDawa, Harari and northern Somali regions. The primary and secondary clusters were located in Dire Dawa, Gambela, Benishangul-Gumuz, western Oromia and southwest Amhara regions, respectively. Women who lived in the primary, secondary and tertiary clusters were more likely to access health services. The Kriging interpolation of women's health service access revealed that there would be good women's health service access in Benishangul-Gumuz, western Amhara, Dire Dawa, eastern Oromia and northern Somali regions. This finding was supported by a similar study done about women's home delivery that states a high proportion of home delivery is found in Amhara, Afar, Tigray, Oromia, and South Nations Nationalities and People's Region,[30] and incomplete maternal continuum care utilisation.[13] Therefore, policy-makers should give priority attention to the areas where women had less likely to access health services in Ethiopia.

In the multilevel mixed effect logistic regression analysis, secondary and higher educational status, rich wealth status, and exposure to media were positively associated, and being a rural resident was negatively associated with women's health service access, respectively.

Women with secondary and higher education were 1.6 and 2 times more likely to access health services. The current evidence was similar to studies done in Ethiopia[17 18] and the Republic of Vanuatu.[31] This might be education's power to enhance women's health-seeking behaviours, educated women actively involved in reading materials and discussions that would enhance their knowledge.[17] Moreover, educated women might give priority attention

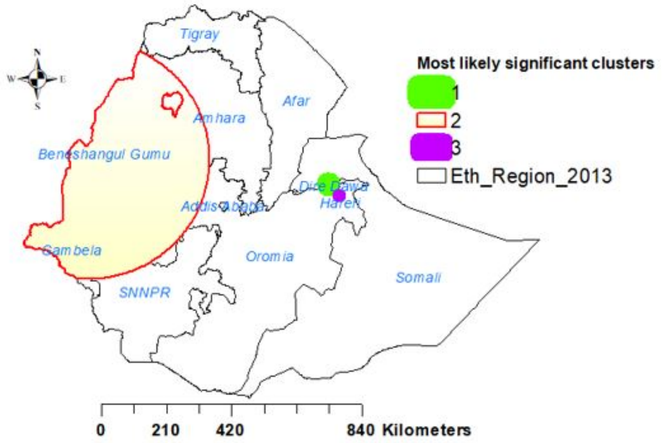


Figure 4 SaTScan analysis of health service access in Ethiopia, 2016 EDHS. EDHS, Ethiopian Demographic and Health Survey.

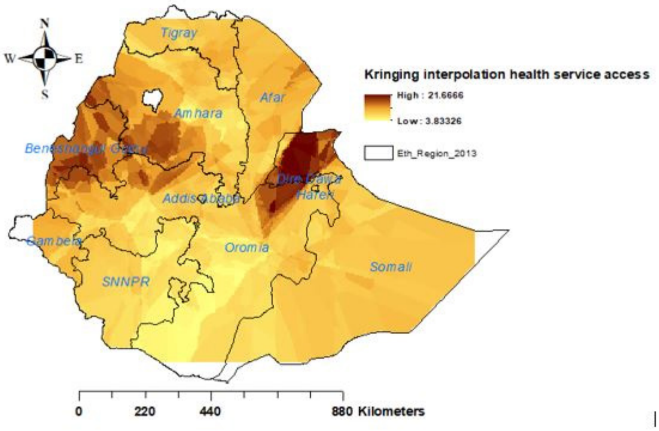


Figure 5 Ordinary kriging interpolation of health service access in Ethiopia, 2016 EDHS. EDHS, Ethiopian Demographic and Health Survey.

Table 3 Multilevel mixed-effect logistic regression analysis of women's health service access using 2016 EDHS data

			Model 2	Model 3	Model 4
Variables	**Category**	**Model 1**	**AOR (95% CI)**	**AOR (95% CI)**	**AOR (95% CI)**
Educational status	Primary		1.19 (1.06 to 1.33)*	-	1.19 (1.07 to 1.34)
	Secondary		1.59 (1.36 to 1.85)*		1.56 (1.34 to 1.81)**
	Higher		2.10 (1.74 to 2.55)*		2.02 (1.66 to 2.44)**
	No education		1		1
Respondent's age	20–24 years		1.04 (0.91 to 1.18)	–	1.03 (0.91 to 1.17)
	25–29 years		1.01 (0.88 to 1.15)	–	1.00 (0.88 to 1.14)
	30–34 years		1.06 (0.92 to 1.23)	–	1.06 (0.92 to 1.23)
	35–39 years		0.96 (0.83 to 1.13)		0.96 (0.83 to 1.12)
	40–44 years		0.95 (0.80 to 1.13)		0.94 (0.79 to 1.12)
	45–49 years		0.86 (0.71 to 1.04)		0.85 (0.70 to 1.03)
	15–19 years		1		1
Wealth status	Rich		1.35 (1.17 to 1.58)*	–	1.38 (1.19 to 1.61)**
	Middle		2.37 (2.04 to 2.74)*	–	2.26 (1.93 to 2.62)
	Poor		1		1
Respondents currently working	Yes		0.98 (0.91 to 1.09)		0.96 (0.90 to 1.09)
	No		1		1
Media exposure	Yes		1.16 (1.04 to 1.30)*	–	1.15 (1.03 to 1.29)**
	No		1		1
Region	Afar		–	0.64 (0.42 to 0.97)*	0.97 (0.65 to 1.45)
	Amhara		–	1.49 (1.03 to 2.17)	1.56 (1.10 to 2.22)
	Oromia		–	0.30 (0.21 to 0.45)	0.28 (0.19 to 0.40)
	Somali		–	0.58 (0.40 to 0.85)*	0.92 (0.64 to 1.35)
	Benishangul		–	0.52 (0.34 to 0.80)	0.56 (0.38 to 0.83)
	SNNPR		–	0.56 (0.38 to 0.82)*	0.53 (0.37 to 0.76)
	Gambela		–	0.67 (0.44 to 1.01)	0.78 (0.52 to 1.14)
	Harari		–	2.77 (1.79 to 4.29)	2.51 (1.66 to 3.79)
	Addis Abeba		–	0.89 (0.59 to 1.37)	0.66 (0.44 to 0.98)
	Dire Dawa		–	0.25 (0.16 to 0.39)*	0.25 (0.16 to 0.38)
	Tigray			1	1
Residency	Rural	–	–	0.19 (0.15 to 0.24)*	0.18 (0.14 to 0.23)**
	Urban			1	1
Model comparison	ICC	0.40	0.28	0.23	0.19
	Variation	0.17	0.105	0.081	0.072
	MOR (95% CI)	2.22 (1.92 to 2.57)	1.29 (1.10 to 1.52)	0.97 (0.82 to 1.15)	0.81 (0.68 to 0.97)
	AIC	16866	16218	16502	16027

*Significant at model 2 and model 3; **, significant at model 4.
AIC, Akaike's information criteria; AOR, adjusted odds ratio; CI, confidence interval; EDHS, Ethiopian demographic and health survey; ICC, intraclass correlation coefficient; MOR, median odds ratio; SNNPR, South Nations Nationalities and People's Region.

to their health, strive to know the benefits of healthcare services and illiterate women may fail to receive health services during pregnancy.[32] In line with this finding, stakeholders should enhance women's educational status by using different educational delivery mechanisms, for instance, a health professional could provide appropriate consultation service during women's health facility visits, and educational messages for women could be sent to women through short message services. Rich women were 1.4 times more likely to access health services. This finding was similar to studies done in Ethiopia[17 18] and the Republic of Vanuatu.[31] This might be women's better economic status which increases their healthcare-seeking behaviour and autonomy in healthcare decision-making, they may afford to cover medical and transportation costs. Furthermore, wealthy women may cover their drug and transportation costs.[18] In addition, poor women could have poor utilisation of preventive, promotive and curative aspects of health services. So, policy-makers should enhance women's wealth status, and encourage women to have their daily income.

The women who had media exposure were 1.2 times more likely to access health services. This finding was similar to studies done in Ethiopia[13 18] and Nepal.[33] This could be the power of mass media in disseminating information concerning maternal health that may enhance women's knowledge and attitude towards health service access and utilisation.[33] Furthermore, women who were exposed to the media were more likely to be informed about health services utilisation. Therefore, the availability of media spots is critical to the delivery of health-related information messages that could reach out to women in their homes.

Rural resident women were 82% less likely to access health services. This finding was similar to studies done in Ethiopia.[17 18] This might be because health facilities are inadequately accessible and available in rural areas. Rural resident women might be limited in access to education and health information.[27] Moreover, in rural areas adequate health professionals might not be available and so appropriate counselling services might not be delivered. In resource-limited settings, health facilities and necessary infrastructure such as roads and clean water are less likely available in the rural side of the country. Therefore, stakeholders better if they close such gaps in the rural areas to ensure women's health service access and utilisation.

CONCLUSIONS

In Ethiopia, inadequate numbers of women had good health service access. Women had faced problems with getting money for treatment, distance to health facilities, not wanting to go alone to health facilities and getting permission to go for medical care. The distribution of women's health service access was spatially clustered in Ethiopia. Women's health services access was positively associated with education, wealth and media exposure. However, rural resident women were negatively correlated with health service access. Therefore, stakeholders should pay priority attention to the cold spot areas of women's health service access. Improving women's educational status, and providing women with various media access, is critical for health service access. In addition, stakeholders should create job opportunities for poor women, and deliver public health promotion regarding maternal and child health service utilisation. Providing appropriate consolation services and constructing nearby health facilities are also possible interventions to enhance women's health service access.

Strengths and limitations

This study analysed nationally representative data and a multilevel logistic regression analysis model that could alleviate data correlations were employed. Appropriate health intervention techniques that would increase women's health services access were highlighted. Since the study was based on cross-sectional, social desirability and recall bias might exist. So, the finding might have a temporal relationship. Moreover, the coordinate file was not originally collected at the four-corner direction of Ethiopia. So, this study excludes areas that had no coordination file (at four corners of Ethiopia), and irregularly shaped clusters that were not detected in the SaTScan analysis were excluded.

Acknowledgements We would like to express our deepest appreciation to the DHS programme for permitting data access.

Contributors AWD had substantial contributions to the study design, conception, data analysis, result interpretation and discussion. AWD and ADW had significant contributions to drafting the manuscript. Both authors read and approved the final manuscript for submission for publication.

Funding No funding was received for this study.

Competing interests The author declared that they do not have competing interests.

Patient consent for publication Not applicable.

Ethics approval This study was based on publicly available secondary data source. So, it does not need any ethical clearance. So, it does not need any informed consents from the study participants.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement All data relevant to the study are included in the article or uploaded as online supplemental information.

ORCID iDs
Addisalem Workie Demsash http://orcid.org/0000-0002-9356-8126
Agmasie Damtew Walle http://orcid.org/0000-0001-9583-5876

REFERENCES

1 Tarekegn SM, Lieberman LS, Giedraitis V. Determinants of maternal health service utilization in Ethiopia: analysis of the 2011 Ethiopian Demographic and Health Survey. *BMC Pregnancy Childbirth* 2014;14:161.
2 World Health Organization. Maternal mortality and child health fact sheet. 2012. Available: https://www.who.int/en/news-room/fact-sheets/detail/maternal-mortality
3 World Health Organization. *Trends in maternal mortality: 1990-2015: estimates from WHO, UNICEF, UNFPA, World Bank Group and the United nations population division*. World Health Organization, 2015.
4 Mekonnen Y, Mekonnen A. Factors influencing the use of maternal healthcare services in Ethiopia. *J Health Popul Nutr* 2003;21:374–82.
5 Tegegne TK, Chojenta C, Getachew T, *et al*. Antenatal care use in Ethiopia: a spatial and multilevel analysis. *BMC Pregnancy Childbirth* 2019;19:399.
6 DHS Program. Ethiopia demographic and health survey 2016. 2016.
7 Belay DG, Aragaw FM, Anley DT, *et al*. Spatiotemporal distribution and determinants of delayed first antenatal care visit among reproductive age women in Ethiopia: a spatial and multilevel analysis. *BMC Public Health* 2021;21:1570.
8 Sisay D, Ewune HA, Muche T, *et al*. Spatial distribution and associated factors of institutional delivery among reproductive-age women in Ethiopia: the case of Ethiopia demographic and health survey. *Obstet Gynecol Int* 2022;2022:4480568.
9 Sisay MM, Geremew TT, Demlie YW, *et al*. Spatial patterns and determinants of postnatal care use in Ethiopia: findings from the 2016 demographic and health survey. *BMJ Open* 2019;9:e025066.
10 Akunga D, Menya D, Kabue M. Determinants of postnatal care use in Kenya. *Apst* 2014;28:1447–59.

11 Dahiru T, Oche OM. Determinants of antenatal care, institutional delivery and postnatal care services utilization in Nigeria. *Pan Afr Med J* 2015;21:321.
12 Tesfaye G, Loxton D, Chojenta C, *et al*. Delayed initiation of antenatal care and associated factors in Ethiopia: a systematic review and meta-analysis. *Reprod Health* 2017;14:150.
13 Alamneh TS, Teshale AB, Yeshaw Y, *et al*. Barriers for health care access affects maternal continuum of care utilization in Ethiopia; spatial analysis and generalized estimating equation. *PLoS One* 2022;17:e0266490.
14 Banda C. *Barriers to utilization of focused antenatal care among pregnant women in Ntchisi district in Malawi*. 2013.
15 Baffour-Awuah A, Mwini-Nyaledzigbor PP, Richter S. Enhancing focused antenatal care in Ghana: an exploration into perceptions of practicing midwives. *International Journal of Africa Nursing Sciences* 2015;2:59–64.
16 Alkema L, Chou D, Hogan D, *et al*. Global, regional, and national levels and trends in maternal mortality between 1990 and 2015, with scenario-based projections to 2030: a systematic analysis by the un maternal mortality estimation inter-agency group. *The Lancet* 2016;387:462–74.
17 Yeneneh A, Alemu K, Dadi AF, *et al*. Spatial distribution of antenatal care utilization and associated factors in Ethiopia: evidence from Ethiopian demographic health surveys. *BMC Pregnancy Childbirth* 2018;18:242.
18 Tesema GA, Mekonnen TH, Teshale AB, *et al*. Individual and community-level determinants, and spatial distribution of institutional delivery in Ethiopia, 2016: spatial and multilevel analysis. *PLoS ONE* 2020;15:e0242242.
19 Assefa Y, Damme WV, Williams OD, *et al*. Successes and challenges of the millennium development goals in Ethiopia: lessons for the sustainable development goals. *BMJ Glob Health* 2017;2:e000318.
20 WHO. *Standards for improving quality of maternal and newborn care in health facilities*. 2016.
21 Demsash AW, Tegegne MD, Wubante SM, *et al*. Spatial and multilevel analysis of sanitation service access and related factors among households in Ethiopia: using 2019 Ethiopian national dataset. *PLOS Glob Public Health* 2023;3:e0001752.
22 The 2016 Ethiopian demography and health survey. n.d. Available: https://dhsprogram.com/methodology/survey/survey-display-478.cfm
23 Demsash AW, Chereka AA, Kassie SY, *et al*. Spatial distribution of vitamin A rich foods intake and associated factors among children aged 6-23 months in Ethiopia: spatial and multilevel analysis of 2019 Ethiopian mini demographic and health survey. *BMC Nutr* 2022;8:77.
24 Chaikaew N, Tripathi NK, Souris M. Exploring spatial patterns and hotspots of diarrhea in Chiang Mai, Thailand. *Int J Health Geogr* 2009;8:36.
25 Kulldorff M. A spatial scan statistic. *Communications in Statistics - Theory and Methods* 1997;26:1481–96.
26 Alemu K, Worku A, Berhane Y, *et al*. Spatiotemporal clusters of malaria cases at village level, Northwest Ethiopia. *Malar J* 2014;13:223.
27 Babalola S, Fatusi A. Determinants of use of maternal health services in Nigeria -- looking beyond individual and household factors. *BMC Pregnancy Childbirth* 2009;9:43.
28 Kebede A, Hassen K, Nigussie Teklehaymanot A. Factors associated with institutional delivery service utilization in Ethiopia. *Int J Womens Health* 2016;8:463–75.
29 Mekonnen T, Dune T, Perz J, *et al*. Trends and determinants of antenatal care service use in Ethiopia between 2000 and 2016. *Int J Environ Res Public Health* 2019;16:748.
30 Tessema ZT, Tiruneh SA. Spatio-Temporal distribution and associated factors of home delivery in Ethiopia. Further multilevel and spatial analysis of Ethiopian demographic and health surveys 2005-2016. *BMC Pregnancy Childbirth* 2020;20:342.
31 Rahman M, Haque SE, Mostofa MG, *et al*. Wealth inequality and utilization of reproductive health services in the Republic of Vanuatu: insights from the multiple indicator cluster survey, 2007. *Int J Equity Health* 2011;10:58.
32 Haque A, Zulfiqar M. Women's economic empowerment through financial literacy, financial attitude and financial wellbeing. *International Journal of Business and Social Science* 2016;7:78–88.
33 Tamang TM. Factors associated with completion of continuum of care for maternal health in Nepal. IUSSP XXVIII International Population Conference, Cape Town, South Africa; 2017

BMJ Health & Care Informatics

Cluster analysis of dietary patterns associated with colorectal cancer derived from a Moroccan case–control study

Noura Qarmiche ,[1] Khaoula El Kinany,[2] Nada Otmani,[3] Karima El Rhazi,[2] Nour El Houda Chaoui[1]

To cite: Qarmiche N, El Kinany K, Otmani N, *et al*. Cluster analysis of dietary patterns associated with colorectal cancer derived from a Moroccan case–control study. *BMJ Health Care Inform* 2023;**30**:e100710. doi:10.1136/bmjhci-2022-100710

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/bmjhci-2022-100710).

Received 30 November 2022
Accepted 28 March 2023

[1]Laboratory of Artificial Intelligence, Data Science and Emerging Systems, National School of Applied Sciences, Sidi Mohamed Ben Abdellah University, Fes, Morocco
[2]Department of Epidemiology, Clinical Research and Community Health, Sidi Mohamed Ben Abdellah University, Fes, Morocco
[3]Health Informatics and Statistics Unit, Department of Epidemiology, Clinical Research and Community Health, Sidi Mohamed Ben Abdellah University, Fes, Morocco

Correspondence to
Noura Qarmiche;
noura.qarmiche@usmba.ac.ma

ABSTRACT

Introduction Colorectal cancer (CRC) is a global public health problem. There is strong indication that nutrition could be an important component of primary prevention. Dietary patterns are a powerful technique for understanding the relationship between diet and cancer varying across populations.

Objective We used an unsupervised machine learning approach to cluster Moroccan dietary patterns associated with CRC.

Methods The study was conducted based on the reported nutrition of CRC matched cases and controls including 1483 pairs. Baseline dietary intake was measured using a validated food-frequency questionnaire adapted to the Moroccan context. Food items were consolidated into 30 food groups reduced on 6 dimensions by principal component analysis (PCA).

Results K-means method, applied in the PCA-subspace, identified two patterns: 'prudent pattern' (moderate consumption of almost all foods with a slight increase in fruits and vegetables) and a 'dangerous pattern' (vegetable oil, cake, chocolate, cheese, red meat, sugar and butter) with small variation between components and clusters. The student test showed a significant relationship between clusters and all food consumption except poultry. The simple logistic regression test showed that people who belong to the 'dangerous pattern' have a higher risk to develop CRC with an OR 1.59, 95% CI (1.37 to 1.38).

Conclusion The proposed algorithm applied to the CCR Nutrition database identified two dietary profiles associated with CRC: the 'dangerous pattern' and the 'prudent pattern'. The results of this study could contribute to recommendations for CRC preventive diet in the Moroccan population.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Diet and lifestyle are believed to play a significant role in the onset of colorectal cancer (CRC).

WHAT THIS STUDY ADDS

⇒ This study investigates this relationship by analysing dietary patterns in Morocco through the use of K-means clustering in a principal component analysis subspace.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ The results provide a clearer understanding of the link between dietary habits and CRC in Morocco, enabling the creation of tailored recommendations.

INTRODUCTION

Colorectal cancer (CRC) is one of the most malignant cancers and the third-leading cause of cancer death in the word[1] accounting for approximately 700 000 annual deaths worldwide.[2]

Diet and lifestyle are likely to play an important role in the development of CRC, but the complexity of this effect is still unclear. Previous studies have focused on the effects of a single food or nutrient and overlooked the interaction or synergy of foods.[3] Dietary patterns analyses are a broader picture of food and nutrient intake. This is an alternative and complementary approach to exploring the relationship between diet and CRC risk. Thus, in recent years, there has been increasing interest in identifying dietary patterns as consumed by populations.[4] Knowledge of population specific dietary patterns is important to identify groups at risk for underconsumption or overconsumption of particular nutrients and to create dietary pattern-based guidelines, which may be easier to translate into diets for the public for CRC prevention.

Clustering is an unsupervised machine learning approach. It aims to identify a cluster structure characterised by the maximum data similarity inside a cluster and the maximum data dissimilarity between different clusters.[5] The oldest and most popular clustering method is K-means, which is a vector quantisation algorithm that attempts to partition n observations into k non-overlapping clusters represented by their centroids. The centroid of a cluster is usually the average of the points in that cluster. The K-means method was ranked second among the 10 best data mining algorithms and has become a reference for all

Table 1 Percentage of missing data for each variable

Variables	% of missing data
q1, q11, q17, q31, q32	0.03
q6	0.17
q16	0.2
q15	0.4
q24	0.74
q21p1, q22p1	21.58
q23p1, q23p2	21.61

new proposed methods.[6] It has the advantage of being very simple, robust and efficient. It can be used for a wide variety of data types.[7] Principal component analysis (PCA) is a widely used dimension reduction method. It transforms high-dimensional data into lower-dimensional data. Where coherent patterns can be detected more clearly.[8] PCA is the continuous solution of the cluster membership indicators in the K-means clustering method. Indeed, PCA selects the dimensions with the largest variances to find the best low-rank approximation (in L2 norm) of the data through the singular value decomposition.[8]

Primary objective

The main objective of this study was to identify Moroccan dietary patterns associated with CRC using CRC Nutrition dataset, which is a Moroccan multicentre case–control study. For this, we applied k-means clustering method in a reduced subspace defined by the PCA dimension reduction method.

Related works

Several studies have been conducted on dietary patterns and potential CRC risk in different populations. In Portugal, three dietary patterns were identified: 'healthy', 'low milk and dietary fibre intake' and 'Western' using PCA and Ward's method. This study confirmed the higher risk of CRC in subjects with a 'Western' diet and a 'low intake of milk and dietary fibre'.[9] In a Korean population, a PCA was used to identify three dietary patterns (traditional, Western and conservative). Traditional and conservative patterns were inversely associated with CRC risk.[10]

Among middle-aged Americans, PCA identified three main dietary patterns: a fruit and vegetable pattern, a diet food pattern, and a red meat and potato pattern. Dietary patterns characterised by low frequency of meat and potato consumption and frequent consumption of fruits and vegetables and low-fat foods were consistent with a decreased risk of CRC.[11] Three dietary patterns were defined by PCA labelled 'meat-based', 'plant-based' and 'carbohydrate-based' patterns in Uruguay. The highest risk was positively associated with the meat-based model, whereas the plant-based model was strongly protective. The carbohydrate model was only positively associated with colon cancer risk.[12] Among a Japanese population, three dietary patterns were derived from the PCA: 'conservative', 'western' and 'traditional'. The conservative model showed a reduced association of CRC. The Western model showed a significant positive linear trend for colon. There was no apparent association of the traditional Japanese dietary pattern on overall or site-specific risk of CRC.[13] A Canadian population-based study identified three main dietary patterns using factor analysis, namely a meat-based diet pattern, a plant-based diet pattern and a sugar-based diet pattern. The results suggest that the meat-based diet and the sugar-based diet increase the risk of CRC. In contrast, the plant-based diet decreased the risk of CRC.[14]

For most of these studies, data were obtained by case–control surveys and dietary intakes were assessed using the food-frequency questionnaire (FFQ).

MATERIALS AND METHODS

Study design

This was a Moroccan, national, retrospective, non-interventional and multicentre study in patients wityh CRC.

Setting

This study was conducted in five major University Hospital centres in Morocco, namely Hassan II UHC of Fez, Avicenna UHC of Rabat, Mohammed VI UHC of Oujda, Averroes UHC of Casablanca and Mohammed VI UHC of Marrakech between September 2009 and February 2017. Participating centres were distributed across the country to ensure geographical representation.

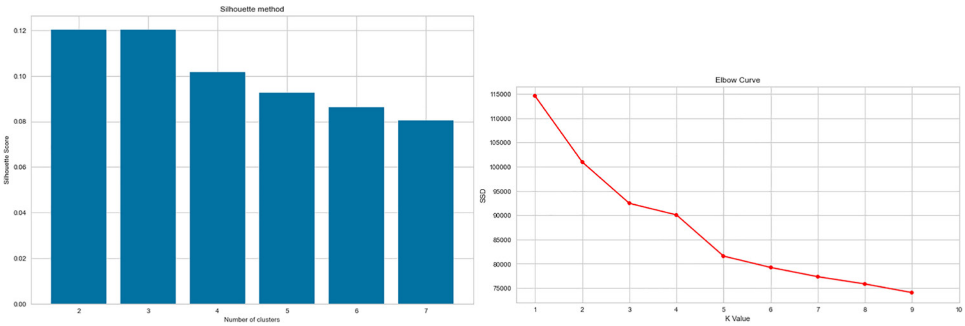


Figure 1 Elbow curve and silhouette histogram.

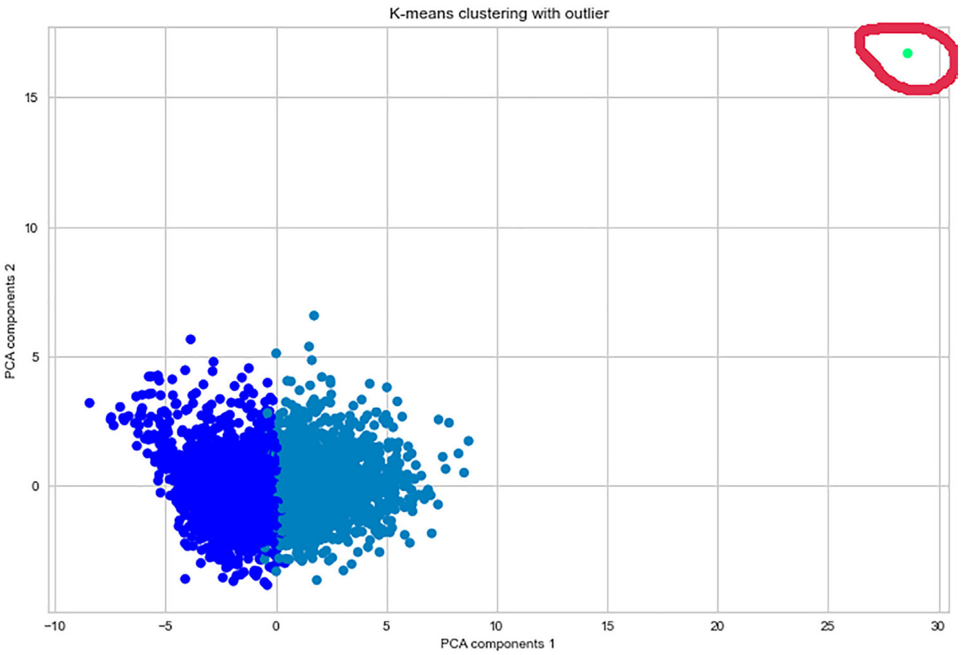


Figure 2 K-means clustering for outlier detection. PCA, principal component analysis.

Participants

Cases and controls were individually matched on age (±5 years), sex and centre (ratio 1:1). Cases were defined as patients who had recently confirmed CRC diagnosis by histopathology and who did not start any treatment protocol (chemotherapy, radiotherapy, hormonal therapy or surgery) at the time of inclusion. Other eligibility criteria were 18 years of age or older, no history of diabetes mellitus, ability to give consent and ability to communicate and conduct the interview. Controls were selected from the same local population and hospitals as the cases, among healthy subjects accompanying other patients or visitors. Cases and controls both met the same eligibility requirements, with the exception of the criterion that did not have a personal history of CRC or any other type of cancer.[10 15]

Data collection

Data were collected in face-to-face interviews conducted by trained interviewers. All participants were invited to answer questions on the following topics: sociodemographic information (age, sex, centre, residency, profession, marital status, education level, income level and type of habitat), clinical data, substances use, physical activity levels, anthropometric measurements, genetic data and dietary data. Dietary information was obtained via a validated semiquantitative FFQ. This questionnaire was based on the GA2LEN FFQ and was adapted to the Moroccan context.[16] To objectively assess the frequency of food consumption, a detailed frequency scale has been established, including the following options: rarely/never, once to three times per month, once/week, twice to four/week, five to six times/week, once/day, twice to three times/day and equal or more than four times/day.[17]

The 255 FFQ items were initially combined into 30 different food and beverage groups, as follows: bread, breakfast with grains, couscous, pasta, cake, rice, sugar, sweets without chocolate, chocolate, vegetable oil, margarine and vegetable fat, butter and animals fat,

Table 2 Total variance explained by the principal components

Principal component	Eigenvalue	% variance	% cumulative
1	4901	16 899	16 899
2	2354	8119	25 017
3	1729	5961	30 978
4	1512	5213	36 191
5	1351	4659	40 850
6	1220	4207	45 057

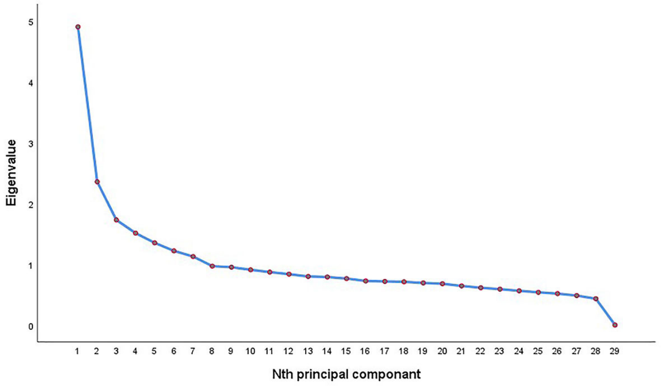


Figure 3 PCA scree plot. PCA, principal component analysis.

Table 3 Principal component loadings (correlations between features and principal components (r-value))

CP1	CP2	CP3	CP4	CP5	CP6
Vegetable oil (0.85) Cake (0.65) Chocolate (0,58) Miscellaneous_ foods (0.54) Milk (0.53) Nuts(0.5) Juice(0.49) Rice(0.44) Sugar(0.44) Other dairy products (0.43) Cheese (0.42) Non-alcoholic_ beverages (0.41) pasta (0.41)	Vegetables (0.47) Red meat (–0.45) Breakfast with_ grains (0.42) Offal (–0.42)	Sweets except chocolate (–0.52)	Fruits (–0.52) Fish (–0.5)	Poultry (0.55) Potatoes (0.41)	Bread (–0.45)

nuts, legumes, vegetables, potatoes, fruits, juice, non-alcoholic beverages, coffee/tea, meat, dried meat, poultry, offal, fish, milk of cow/milk of soya, cheese, other dairy products, miscellaneous foods and alcohol. The details of the components of each group are detailed here.[16]

Bias

This non-interventional study is subject to various biases and structural limitations inherent in observational studies. Participants recorded their usual food intake over a longer period (1 year), which could lead to errors in the results. This information bias was addressed at the time of recruitment by trained investigators who collected the data with maximum accuracy. To account for potential confounders in this study, a large amount of data that could affect exposure and outcomes (such as physical activity, body mass index (BMI), alcohol and tobacco use) were collected, and the data were fairly complete for the outcomes.

Study size

The sample size for the study was determined by taking into account the prevalence of red meat consumption as a key exposure of interest. Data from the National Survey of Dietary Habits in Morocco revealed that 62.7% of Moroccan adults eat red meat at least twice a week. The following formula specific for individual-matched case–control studies, the sample size was calculated with 5% type I error, a 90% statistical power and a minimum difference in risk of 43% as reported by the WCRF/AICR report.

$$n_c = \frac{\left(Z_\alpha(\psi+1)+2Z_\beta\sqrt{\psi}\right)}{(\psi-1)^2(\psi+1)P_{01}}$$

Where n_c= sample size for case–control pairs.

ψ= OR.

P_0= The probability of obtaining a matched pair in which the case is unexposed and the control is exposed.

The number of pairs needed for the study was 1496 rounded to 1500.

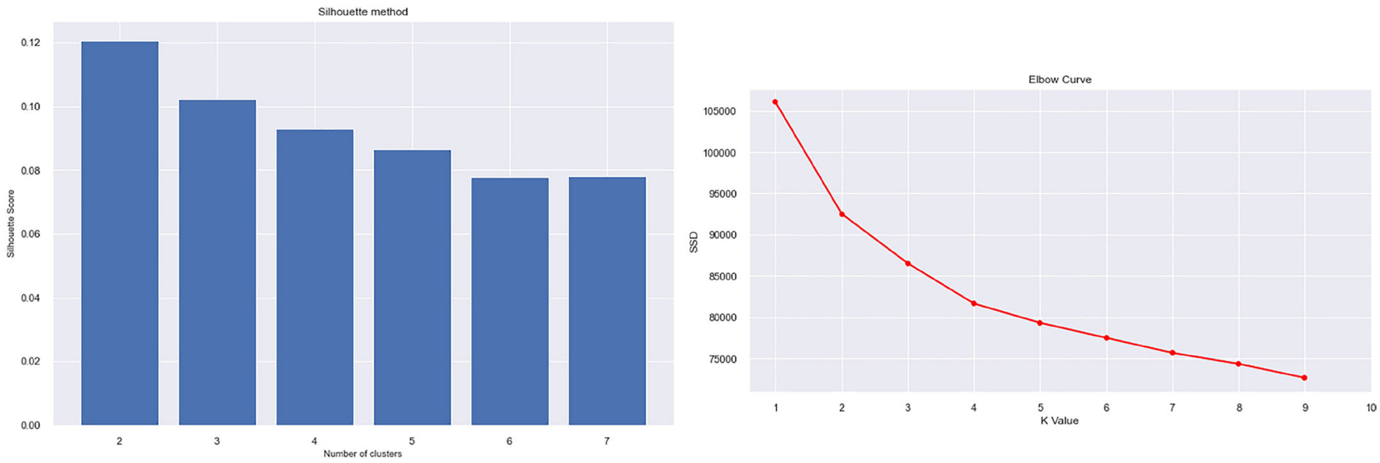


Figure 4 Elbow curve and silhouette histogram after outlier removal. SSD, sum of squares of distances

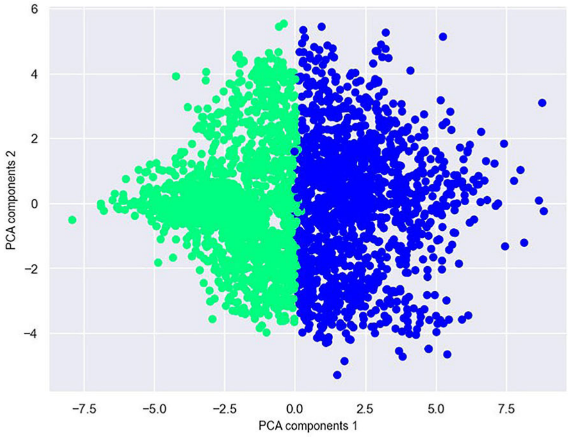


Figure 5 K-means clustering after outlier removal. PCA, principal component analysis.

Statistical analyses

Data cleaning and handling

In total, 3032 participants were recruited for the study, 1516 cases and 1516 controls. However, 7 participants with unspecified primary cancer, 6 cases with old biopsies, 10 participants with missing dietary data, 2 duplicate records and 8 unmatched records were excluded.

The participation rate in this study was 97% (1516/1555) for cases and 76% (1516/2000) for controls. The final sample included in this study was 1483 cases and 1483 controls.

Data preprocessing

Missing values for each variable were replaced by its mean if the percentage of missing data for that variable is less than 20%, otherwise the variable will be removed from the study.[18] SimpleImputer, which is a sklearn class, was used as imputation method.

All FFQ values are on the same scale and are between 2 and 9, so there was no need to normalise them.

K-means method has been used to detect outliers, which are extreme values, abnormally different from the variable distribution.[19] In clustering analyses, they are in the form of too small groups that must be removed.[20] Detecting outliers allows improving the quality of clustering.[21]

Unsupervised learning algorithms

Principal component analysis

PCA, a dimensionality reduction algorithm, was used to reduce the number of food groups by mapping each instance of a given data set to a k-dimensional subspace called principal components, where k<d. The scree plot was used to identify the number of principal components to retain, which shows the proportion of variance explained by each component. The first component covers most of the model and covers the maximum variance, while each subsequent component covers a lesser value of the variance.[22]

Table 4 Characteristics of consumption across the two dietary patterns

Cluster	Prudent pattern (mean±SD)	Dangerous pattern (mean±SD)	P value
CP1			
Oil	2.78±0.42	3.61±0.69	<0.001
Cake	2.49±0.9	4.04±1.52	<0.001
Chocolate	2.28±0.78	3.27±1.5	<0.001
Miscellaneous foods	2.41±0.3	2.62±0.46	<0.001
Milk	2.76±0.44	3.11±0.46	<0.001
Nuts	2.51±0.94	3.3±1.27	<0.001
Juice	2.38±0.61	2.74±0.84	<0.001
Rice	3.12±1.02	3.78±0.96	<0.001
Sugar	3.83±0.77	4.32±0.84	<0.001
Other dairy products	2.04±0.21	2.13±0.4	<0.001
Cheese	3.37±1.74	5.39±1.57	<0.001
Non-alcoholic_ beverages	2.58±0.73	2.84±0.91	<0.001
Pasta	2.89±1.04	3.8±1.08	<0.001
CP2			
Vegetables except potatoes	6.3±1.23	6.2±1.04	<0.001
Red meat	4.21±1.22	4.44±1.15	<0.001
Breakfast with grains	2.79±1.22	3.24±1.39	<0.001
Offal	2.15±0.37	2.24±0.47	<0.001
CP3			
Sweets Except chocolate	2.17±0.72	2.62±1.24	<0.001
CP4			
Fruits	5.33±1.49	5.15±1.28	<0.001
Fish	3.74±0.98	4.19±0.96	<0.001
CP5			
Poultry*	4.48±0.98	4.49±0.95	0.586
Potatoes	5.19±1.41	5.46±1.07	<0.001
CP6			
Bread	8.03±0.97	8.32±0.71	<0.001

K-means clustering

K-means clustering aims to divide M points in N dimensions into a set C of K clusters Cj with cluster mean cj to reduce the sum of squared errors.[23 24] This is described as follows:

$$E = \sum_{j=1}^{k} \sum_{x_i \in c_j} \|c_j - x_i\|^2 \qquad (1)$$

Where, E is sum of the square error of objects with cluster means for K cluster and distance metric between a data point and a cluster mean. The Euclidean distance is defined as:

$$\|x - y\| = \sqrt{\sum_{i=1}^{v} |x_i - y_i|^2} \qquad (2)$$

Table 5 Distributions of sociodemographic characteristics of the study population by the two clusters

	Prudent pattern	Dangerous pattern	P value
Age			
[18–30[	1.96	2.29	0.753
[30–45[	9.68	9.04	
[45–60[	20.17	18.58	
[60–75[	16.12	14.94	
>75	3.61	3.61	
Sex			
Female	25.94%	24.38%	0.994
Mal	25.60%	24.08%	
Marital status			
Single	5.16%	4.55%	0.086
Married	38.48%	37.91%	
Divorced	1.92%	1.65%	
Widowed	5.97%	4.35%	
Residence			
Urban	35.35%	36.73%	<0.001
Rural	16.19%	11.74%	
Level of education			
Illiterate	31.10%	25.67%	
Primary	9.75%	9.04%	0.001
Secondary	7.05%	8.09%	
Higher	3.64%	5.67%	
Profession			
Unemployed	7.82%	5.60%	0.001
Housewife	20.13%	17.17%	
Student	0.30%	0.51%	
Working	19.43%	20.57%	
Retired	3.84%	4.62%	
Smoking status			
Non-smoker	42.12%	40.92%	0.95
Smoker	5.40%	4.75%	
BMI (kg/m^2)			
(16–18.5)	1.00	1.07	
(18.5–25)	21.73	20.87	
(25–30)	21.73	21.25	0.1
≥30	7.13	5.23	
Physical activity			
Yes	10.56%	11.32%	0.061
No	40.98%	37.13%	
Monthly household income (DHMAD)			
≤2000	42.80%	34.00%	0.001
(2000–5000)	6.64%	10.42%	
(5000–10 000)	2.09%	4.05%	

BMI, body mass index.

Following vector defines the average of a cluster by:

$$c_j = \frac{1}{|c_j|}\sum_{i\in c_j} x_i \tag{3}$$

Choice of the optimal number of clusters K

In order to determine the optimal number of clusters, we used the Elbow method complemented by silhouette analysis, which calculates the separation distance between the resulting clusters and provides a way to visually assess their number.[25–27]

Proposed method

The K-means method has been applied in the PCA-subspace, as strongly advised by several studies.[8 28 29] Indeed, the continuous solution of the cluster indicators is given by the PCA principal components and the optimal solution of the K-means clustering is inside the PCA-subspace .

Association test

To test the association between clusters and CRC status, the simple logistic regression test was used. Result was presented by OR value and its CI.

Student's t-test was used to assess the relationship between the clusters and food consumption. P values less than 0.05 were considered statistically significant.

The algorithm proposed in this study is presented in online supplemental figure 1.

RESULTS

Data preprocessing

Managing missing data

The number of missing values was calculated by the isnull().sum() function of Pandas. The results obtained are presented in table 1 (only the variables that contained missing data have been reported).

Missing data for variables q1, q6, q11, q15, q16, q17, q24, q31, q32 were replaced by the mean, using Sklearn's simple imput function.

The variables q21p1, q22p1, q23p1, q23p2 that corresponds to alcohol consumption were removed from the study because they contained more than 20% of missing data.

Detection of outliers

The Elbow and Silhouette methods (figure 1) indicate that the appropriate number of clusters k is 3.

K-means identified three distinct groups in our population study (figure 2.). However, it is very evident that one of the groups is simply an outlier since it contains only one point. After checking the database, we verified the existence of an outlier (q15=99) and deleted the record corresponding to this value before running our algorithm again with the new database.

Dimensionality reduction

According to the scree plot figure 3, we have retained six principal components, which were defined by PCA.

From table 2, we notice that the first principal component constitutes 16.89% of the variance. The composition of the first and second axis constitutes 25.01% of the total variance. While the cumulative variance of the 6 principal components represents 45.05% of the total.

The correlation of each principal component with its constituents is presented in table 3 (only correlations >0.4 are reported).

K-means clustering

The results of the Elbow and Silhouette methods (figure 4) indicate that the appropriate number of clusters k is 2.

K-means clustering identified two distinct groups in this population (figure 5). A total of 1433 participants (48.33%) were in cluster 0 while 1531 (51.67%) were in cluster 1. 55.95% of individuals in cluster 0 were controls while 44.04% were cases. Cluster 1 is composed of 44.41% controls and 55.59% cases.

Mean and SD consumption of food groups in each cluster are shown in table 4. The p value between groups was significant (<0.001) for most food groups, with the exception of poultry (p=0.586).

We describe cluster 1 as a 'dangerous pattern' because it showed high loadings of vegetable oil, cake, chocolate, cheese, red meat, sugar and butter. Cluster 0 was termed the 'prudent diet' cluster due to moderate consumption of almost all foods with a slight increase in fruits and vegetables (online supplemental figure 2).

The student test showed a significant relationship between CRC and cluster (p<0.001). Indeed, people who belong to the 'dangerous pattern' have a higher risk to develop CRC with an OR 1.59 (95% CI 1.375 to 1.383).

The distributions of sociodemographic characteristics by cluster are presented in table 5 . No significant differences between dietary patterns were found by age, sex, BMI, marital status, physical activity or smoking status with p values equal to 0.753, 0.994, 0.1, 0.086, 0.061 and 0.95, respectively.

The proportions of the unemployed and housewives were greater in the conservative profile, while the proportions of working and retired people were higher in the dangerous cluster. We also note that the number of people in the dangerous cluster increases proportionally with income and educational level.

DISCUSSION

The proposed algorithm applied to the CCR Nutrition database, which is a multicente case–control study conducted in a population of 1496 pairs of Moroccan subjects with and without CRC, identified 2 dietary profiles associated with CRC: the 'dangerous pattern' and the 'prudent profile'. The 'dangerous pattern' was characterised by a high consumption of vegetable oil, cakes, chocolate, cheese, red meat, sugar and butter. While the 'prudent pattern' was characterised by a moderate consumption of almost all foods with a slight increase in fruits and vegetables. The frequency of cases was higher in the 'dangerous' group than in the 'prudent' group.

This study proposes a new methodological approach that combined two unsupervised machine-learning techniques: PCA and K-means. The K-means method has been applied in the PCA-subspace. Several studies have shown the advantages of this approach.[8 18 28] Indeed, the continuous solution of the cluster indicators is given by the principal components of the PCA and the optimal solution of the K-means clustering is in the PCA subspace. Moreover, the performance of clustering is better at reduced cost and noise. A recent statistical methods review for dietary pattern analysis reported the advantages and the disadvantages of PCA and k-means clustering algorithm. Compared with traditional statistical methods, classification via machine learning techniques reduces misclassification rate, increases generalisability, allows grading of movement quality, and simplifies experimental design.

Other strengths of our research should be mentioned; first, it is the first study on the clustering of dietary profiles related to CRC in Morocco by an unsupervised machine learning approach, according to the literature search. On the other hand, in our case–control study, we included recent diagnosed CRC cases to avoid diet changes. In addition, trained interviewers ensured FFQ questionnaires fulfilment in order to maintain the responses objectivity.[15]

Two limitations of our study must be highlighted; the first one, our clustering was based on food groups containing foods known to be protective against CRC and others known to be risk factors. Thus, clustering of these foods may neutralise their effects and make discrimination difficult. The second one, food consumption was based on frequencies without considering the daily quantities which can influence the clustering.

A recent study used Global Dietary database (Canada, India, Italy, South Korea, Mexico, Sweden and the USA) found that CRC could be predicted based on a list of important dietary data using supervised and unsupervised machine learning approaches. This study identified the following two patterns, total fat, mono unsaturated fats, linoleic acid, cholesterol, omega-6 as moderate to high correlated dietary features to positive CRC, and fibre and carbohydrates as negative correlation with CRC cases. A systematic review of 17 years of evidence (2010–2016) revealed two distinct global dietary patterns related to CRC risk: a 'healthy' pattern, characterised by high intake of fruits and vegetables, higher intakes of one or more of the following foods; whole grains, nuts and legumes, fish and other seafood, milk and other dairy products, and an 'unhealthy' dietary pattern characterised by high intakes of red and processed meat, sugar-sweetened beverages, refined grains and desserts and potatoes.

Several studies in American, European and Asian populations have found three dietary patterns related to CRC[9 11 13 14 30]: 'Western or meat-based diet' which is related with higher risk of CRC, 'healthy or conservative or prudent' which is related with low risk of CRC and 'low milk and dietary fibre intake or traditional' which is relatively related with higher risk of CRC. We could not obtain a very clear group due to diverse nature of nutrition landscape in the Moroccan population, although there were higher intakes of some harmful foods in the cases compared with the controls (meat, sugar and chocolate). The difference in poultry consumption was non-significant between the two clusters, which was similarly reported in a previous study.[31]

The perspectives of this work are as follows: first to repeat the clustering process, but this time with single foods to overcome the limitation of grouping protective and risk foods in the same group, and neutralise their effect. Second, to develop an easy and user-friendly web application that allows the simple user to identify him/herself in a dietary pattern and evaluate whether he/she is following a healthy diet or not, which is the best approach to make a personal prevention as recommended by the latest WHO guidelines.[32]

CONCLUSION

The combination of the two unsupervised learning methods PCA and K-means identified two clusters describing two main dietary patterns related to CRC in the Moroccan population, labelled: 'prudent' and 'dangerous'. The number of cases was relatively higher in the 'dangerous' group than in the 'prudent' group. The unsupervised learning approach proposed in this paper was effective and confirmed the results of the literature but in a more discriminant manner.

Acknowledgements Many thanks to Lalla Salma Foundation, Prevention and Treatment of Cancers (FLSC) and Moroccan Society of Diseases of the Digestive System (SMMAD) for the financing 'CCR Nutrition' study. Many thanks also to all contributors in the five University Hospitals centres; the directors of UHCs: Fez (Pr. Ait Taleb K), Casablanca (Pr. Afif My H); Rabat (Pr. Chefchaouni Al Mountacer C); Oujda (Pr Daoudi A); and Marrakech (Pr. Nejmi H). The heads of medical services and their teams: Casablanca (Pr. Benider A; Pr Alaoui R; Pr. Hliwa W; Pr. Badre W, Pr. Bendahou K, Pr. Karkouri M.), Rabat (Pr. Ahallat M; Pr. Errabih I; Pr. El Feydi AE; Pr. Chad B; Pr. Belkouchi A; Pr. Errihani H; Pr. Mrabti H; Pr. Znati K), Fez (Pr. Nejjari C; Pr Ibrahimi SA; Pr. El Abkari M; Pr. Mellas N; Pr. Chbani L; Pr. Benjelloun MC), Oujda (Pr. Ismaili N; Pr. Chraïbi M; Pr. Abda N, Pr. Abbaoui S) and Marrakech (Pr. Khouchani M; Pr. Samlani Z; Pr. Belbaraka R; Pr. Amine M)

Contributors KER is the principal investigator of the CCR Nutrition study and participated in the writing of the document. KEK collected the data and extracted the dietary data from the database. NO participated in statistical analyses and revision of the manuscript. NEHC validated the methodology and verified the writing of the paper. NQ proposed the algorithm, programmed it, wrote the manuscript and she is the author responsible for the overall content as the guarantor

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval This study was approved by the ethics committee of the Hassan II University Hospital in Fes, Morocco. Participants gave informed consent to participate in the study before taking part.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement No data are available.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

ORCID iD

Noura Qarmiche http://orcid.org/0000-0002-1786-5049

REFERENCES

1 900-world-fact-sheets.pdf. 2021. Available: https://gco.iarc.fr/today/data/factsheets/populations/900-world-fact-sheets.pdf
2 Torre LA, Bray F, Siegel RL, *et al*. Global cancer statistics, 2012. *CA Cancer J Clin* 2015;65:87–108.
3 Clinton SK, Giovannucci EL, Hursting SD. The world cancer research fund/american institute for cancer research third expert report on diet, nutrition, physical activity, and cancer: impact and future directions. *J Nutr* 2020;150:663–71.
4 Schwerin HS, Stanton JL, Smith JL, *et al*. Food, eating habits, and health: a further examination of the relationship between food eating patterns and nutritional health. *Am J Clin Nutr* 1982;35(5 Suppl):1319–25.
5 Sinaga KP, Yang MS. Unsupervised K-means clustering algorithm. *IEEE Access* 2020;8:80716–27.
6 Wu X, Kumar V, Ross Quinlan J, *et al*. Top 10 algorithms in data mining. *Knowl Inf Syst* 2008;14:1–37.
7 Wu J. Advances in k-means clustering. In: *Advances in K-means Clustering: A Data Mining Thinking*. Berlin, Heidelberg: Springer Science & Business Media, 2012.
8 Ding C, He X. *K* -means clustering via principal component analysis . Twenty-first international conference; Banff, Alberta, Canada.New York, New York, USA, 2004:29
9 Magalhães B, Bastos J, Lunet N. Dietary patterns and colorectal cancer: a case-control study from Portugal. *Eur J Cancer Prev* 2011;20:389–95.
10 Park Y, Lee J, Oh JH, *et al*. Dietary patterns and colorectal cancer risk in a Korean population: a case-control study. *Medicine (Baltimore)* 2016;95:e3759.
11 Flood A, Rastogi T, Wirfält E, *et al*. Dietary patterns as identified by factor analysis and colorectal cancer among middle-aged Americans. *Am J Clin Nutr* 2008;88:176–84.
12 De Stefani E, Ronco AL, Boffetta P, *et al*. Nutrient-derived dietary patterns and risk of colorectal cancer: a factor analysis in Uruguay. *Asian Pac J Cancer Prev* 2012;13:231–5.
13 Shin S, Saito E, Sawada N, *et al*. Dietary patterns and colorectal cancer risk in middle-aged adults: a large population-based prospective cohort study. *Clin Nutr* 2018;37:1019–26.
14 Chen Z, Wang PP, Woodrow J, *et al*. Dietary patterns and colorectal cancer: results from a Canadian population-based study. *Nutr J* 2015;14:8.
15 Mint Sidi Ould Deoula M, Huybrechts I, El Kinany K, *et al*. Behavioral, nutritional, and genetic risk factors of colorectal cancers in Morocco: protocol for a multicenter case-control study. *JMIR Res Protoc* 2020;9:e13998.
16 El Kinany K, Garcia-Larsen V, Khalis M, *et al*. Adaptation and validation of a food frequency questionnaire (FFQ) to assess dietary intake in Moroccan adults. *Nutr J* 2018;17:61.
17 El Kinany K, Mint Sidi Deoula M, Hatime Z, *et al*. Consumption of modern and traditional Moroccan dairy products and colorectal cancer risk: a large case control study. *Eur J Nutr* 2020;59:953–63.
18 Cismondi F, Fialho AS, Vieira SM, *et al*. Missing data in medical databases: impute, delete or classify? *Artif Intell Med* 2013;58:63–72.
19 Benzaki Y. Tout savoir sur les valeurs aberrantes (outliers). mr. mint: apprendre le machine learning de A à Z. 2017. Available: https://mrmint.fr/outliers-machine-learning

20 DataTechNotes. Anomaly detection example with K-means in python. 2022. Available: https://www.datatechnotes.com/2020/05/anomaly-detection-with-kmeans-in-python.html
21 Dino L. Outlier detection using K-means clustering in python medium. 2022. Available: https://towardsdev.com/outlier-detection-using-k-means-clustering-in-python-214188fc90e8
22 Keerthi Vasan K, Surendiran B. Dimensionality reduction using principal component analysis for network intrusion detection. *Perspectives in Science* 2016;8:510–2.
23 Farhang Y. n.d. Face extraction from image based on k-means clustering algorithms. *Ijacsa*;8.
24 Hartigan JA, Wong MA. Algorithm as 136: a k-means clustering algorithm. *Applied Statistics* 1979;28:100.
25 Umargono E, Suseno JE, Vincensius Gunawan SK. K-means clustering optimization using the elbow method and early centroid determination based on mean and median formula. The 2nd International Seminar on Science and Technology (ISSTEC 2019); Yogyakarta, Indonesia.Paris, France, November 3, 2020:121–9
26 Tout ce que vous voulez savoir sur l'algorithme K-Means. Mr. mint: apprendre le machine learning de A à Z. 2018. Available: https://mrmint.fr/algorithme-k-means
27 Zhou HB, Gao JT. Automatic method for determining cluster number based on silhouette coefficient. *AMR* 2014;951:227–30.
28 Xu Q, Ding C, Liu J, *et al*. PCA-guided search for k-means. *Pattern Recognition Letters* 2015;54:50–5.
29 Refining initial points for K-means clustering | bibsonomy. 2022. Available: https://www.bibsonomy.org/bibtex/29433d748d0d60d70afdeb54f9418baad/ans
30 Garcia-Larsen V, Morton V, Norat T, *et al*. Dietary patterns derived from principal component analysis (PCA) and risk of colorectal cancer: a systematic review and meta-analysis. *Eur J Clin Nutr* 2019;73:366–86.
31 The VARCLUS procedure. In: 43. n.d:
32 Ward JH. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association* 1963;58:236–44.

BMJ Health & Care Informatics

Moving from non-emergency bleeps and long-range pagers to a hospital-wide, EHR-integrated secure messaging system: an implementer report

Ari Ercole, Claire Tolliday, William Gelson, James H F Rudd, Ewen Cameron, Afzal Chaudhry, Fiona Hamer, Justin Davies

To cite: Ercole A, Tolliday C, Gelson W, *et al*. Moving from non-emergency bleeps and long-range pagers to a hospital-wide, EHR-integrated secure messaging system: an implementer report. *BMJ Health Care Inform* 2023;**30**:e100706. doi:10.1136/bmjhci-2022-100706

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/bmjhci-2022-100706).

Received 21 November 2022
Accepted 12 April 2023

Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK

Correspondence to
Dr Ari Ercole; ae105@cam.ac.uk

ABSTRACT

Introduction Obsolete bleep/long-range pager equipment remains firmly embedded in the National Health Service (NHS).

Objective To introduce a secure, chart-integrated messaging system (Epic Secure Chat) in a large NHS tertiary referral centre to replace non-emergency bleeps/long-range pagers.

Methods The system was socialised in the months before go-live. Operational readiness was overseen by an implementation group with stakeholder engagement. Cutover was accompanied by a week of Secure Chat and bleeps running in parallel.

Results Engagement due to socialisation was high with usage stabilising approximately 3 months after go-live. Contact centre internal call activity fell significantly after go-live. No significant patient safety concerns were reported.

Discussion Uptake was excellent with substantial utilisation well before cutover indirectly supporting high levels of engagement. The majority of those who previously carried bleeps were content to use personal devices for messaging because of user convenience after reassurance about privacy.

Conclusion An integrated secure messaging system can replace non-emergency bleeps with beneficial impact on service.

INTRODUCTION

In 2019, the UK Health and Social Care Secretary announced that the National Health Service (NHS) should remove bleeps and pagers for non-emergency communication by the end of 2021.[1 2] While this technology is now in costly obsolescence and pilot studies have shown efficiency saving[3] using smartphone messaging, legacy equipment remains firmly embedded in the NHS. Optimal strategies for adoption have received little attention[4] and barriers to adoption have been identified.[5]

Cambridge University Hospitals (CUH) NHS Foundation Trust has used a comprehensive Electronic Health Record (EHR, Epic Systems Corporation, Verona, Wisconsin, USA) since 2014. An information-governance compliant messaging solution (Epic Secure Chat) allows for messaging from smartphones, tablets or from within the EHR itself (desktop). The system is fully integrated with the patient chart so that messages and all read/reply times become part of the patient record. Large-scale implementation of an EHR-integrated messaging system to replace non-emergency bleeps/long-range pagers in an NHS organisation has not been previously described.

Setting

CUH is a large, tertiary referral centre in the East of England. It offers a diverse range of services with over 1100 beds and approximately 16000 staff. A significant EHR upgrade (from Epic 2017 to the November 2020 version) was undertaken during the implementation period bringing additional Secure Chat functionality. The implementation period also coincided with a major Wi-Fi infrastructure upgrade to give full coverage across the estate.

Our aim was to replace all bleeps/pagers apart from 'cardiac arrest', 'major trauma' and 'fire' with Secure Chat (online supplemental S1).

METHODS

Secure Chat was made available at our organisation in July 2021. A go-live date in early 2022 was initially chosen due to ongoing COVID-19 pandemic disruption and to leverage additional necessary Secure Chat functionality that would only become available after an Epic version upgrade planned for November 2021.

An implementation group with executive responsibility was formed with representation from the hospital's divisional structure to oversee the project. Socialisation was achieved by a network of 'clinical champions' and through regular communications including trust bulletin items, face-to-face and online question and answer events as well as information on screensavers and posters and offering at-the-elbow support in clinical settings. An etiquette guide was published to define appropriate use of different methods of communication. Our safety surveillance is described in (online supplemental S5).

Contact centre (online supplemental S2) workload was a key concern at the time of cutover since any communications difficulties would likely result in a call to an agent for help. For safety a transition period where contact centre operatives would send messages both to Secure Chat and to existing bleeps for 1 week post go-live was planned. Secure Chat would not be available during (un)planned Epic outages for which the contingency was to fall back on an internal directory of alternative contacts securely maintained by the contact centre and this was widely publicised.

Secure Chat allows for various groups to enable team and role-based messaging. Because of system limitations at the time of the original implementation, our hospital had not fully implemented a sign-in system which we could leverage for automatic group creation. Instead, we created 'opt-in' groups to replicate existing roles, relying on staff to opt-in (out) at the beginning (end) of their duties (online supplemental S3).

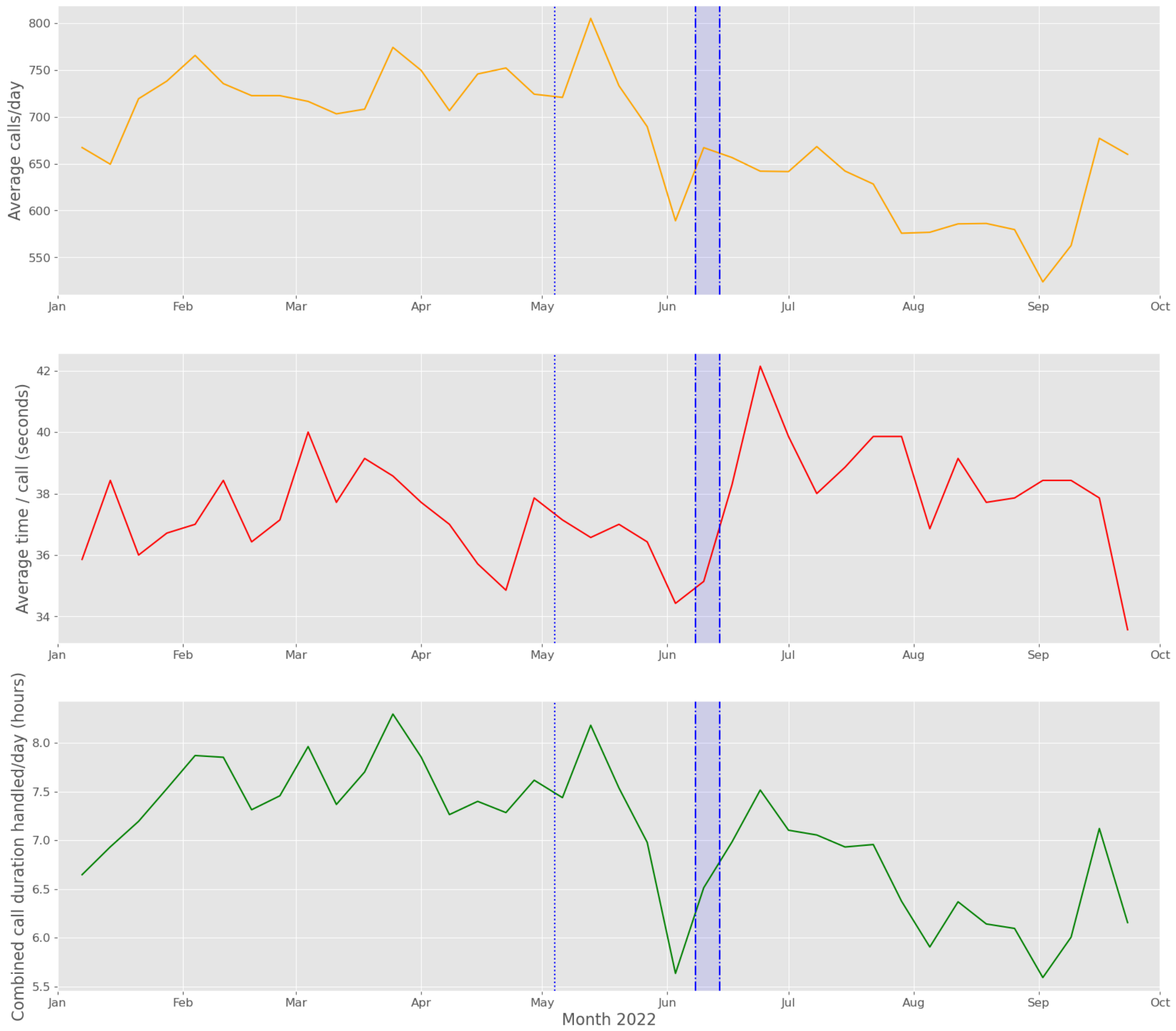


Figure 1 Uptake and organisational impact of Secure Chat implementation. Top panel: internal calls handled per day. Middle panel: average contact centre time spent per call (seconds) Bottom panel: total call time (hours). Dotted line represents date of delayed initial go-live. Data are averaged by week to remove fluctuations from weekends.

Mean comparison was with *t*-tests; structural breaks were examined using the Chow test. Statistical significance was taken at $p<0.05$.

RESULTS

Technical

Secure Chat access was enabled for all members of staff with an Epic login. Non-clinical users, (such as contact centre agents, were not given security to view patient charts). Users who would previously have used bleeps were strongly encouraged to use their own personal devices although mobile phones (or pool phones) were provided in a relatively small number of cases where staff did not have a suitable device or were unwilling.

Workflow

For the mobile app, onboarding involved installation of a CUH-specific profile and a website was set up for this. An initial manual batch activation step was subsequently automated using Blue Prism robotic process automation software (Blue Prism group, Warrington, UK) so that registrations could be completed day and night.

The creation of opt-in groups was a major undertaking and had to be done centrally as no reliable list of baton bleep roles existed. An initial list of some 220 groups was compiled from information from clinical champions and existing bleep lists. After some local user acceptance testing, these groups were made available in December 2021. Inevitably creation, editing and deletion of groups was necessary, and this needed to be done centrally: a review process was set up to ensure consistency.

Outcomes

Adoption through socialisation in the months before go-live across all staff groups was rapid (online supplemental figure S2,S3) across all staff groups with pharmacy (and pharmacy technicians) proving to be an unexpected early adopter. The original 4 May 2022 go-live date was pushed back at a final go/no-go meeting to 8 June 2022 due to isolated specialty-specific readiness concerns. Gross total messages sent plateaued at over 600 000 by 3 months after cutover. Opt-in group maintenance peaked before go-live (online supplemental figure S4) although a significant maintenance burden occurred after the original 4 May date.

Internal call data handled by contact centre operatives is shown in figure 1. The average number of internal calls handled by contact centre operatives fell from 720 to 614 per day ($p<0.0001$) after implementation. While average time/call increased marginally from 37 s to 38 s ($p=0.014$), the total call duration per day fell overall by nearly an hour from 7.4 hours to 6.5 hours per day ($p<0.0001$). There was evidence of significant structural breaks for call numbers and average call time, but not for overall call time ($p=0.01$, 0.0003 and 0.06 respectively).

No significant risk events attributable to the Secure Chat implementation were reported (online supplemental S5).

DISCUSSION

We demonstrate that secure messaging can be implemented in a tertiary NHS hospital without significant incident or negatively impacting on contact centre activity. This was possible even without physically retiring the legacy system: bleep counts dropped to negligible levels (online supplemental figure S3) which is important as multiple coexisting communication methods risk overload.[5]

It is anticipated that the bleep system will be decommissioned in due course depending on a future resilience analysis.

While a minority of staff expressed reservations before go-live citing privacy concerns we were able to provide assurances; most were ultimately content to use their personal devices which offered convenience advantages. The largest complaint received from users concerned inappropriate use of Secure Chat for non-urgent messaging. This is a known issue[3] but the etiquette guide which set out clear expectations was key central to empowering staff to challenge inappropriate messaging.

A number of short (1–2 hours) routine Epic upgrade outages have subsequently taken place (scheduled at weekends and night-time) during which time Secure Chat was not available. Concerns that the contact centre could be overwhelmed at these times have not materialised.

CONCLUSIONS

We were able to effectively replace non-emergency bleeps/long-range pagers with a messaging system integrated with the patient chart in a large NHS academic hospital by the soft approach of socialisation before cutover. Discounting the time before our EHR upgrade in November 2021, we were able to do this in 7 months with message numbers and support needs stabilising within approximately 3 months of go-live using existing infrastructure and without significant incident.

Twitter Ari Ercole @ari.ercole

Acknowledgements The authors thank the CUH eHospital team and Management Executive forsupport and engagement with this initiative. The authors also thank Joseph Scott, Richard Wallis, Benjamin Styllianou, Rose Cormie and Olivia Bentham for their help in obtaining implementation data.

Contributors All authors were involved in the implementation of the project. AE wrote up the project with approval from all other authors.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None at the time of implementation however AC will be leaving the role of Director of Digital at Cambridge University Hospitals in February 2023 to take up a post with Epic Systems Corporation. (This is not a conflict of interest as per the COI form but we include it for complete transparency.)

Patient consent for publication Not applicable.

Ethics approval Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data sharing is not applicable as no data sets were generated and/or analysed for this study.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

ORCID iD

Ari Ercole http://orcid.org/0000-0001-8350-8093

REFERENCES

1 Health and social care secretary bans pagers from the NHS. Available: https://www.gov.uk/government/news/health-and-social-care-secretary-bans-pagers-from-the-nhs [Accessed 25 Sep 2022].

2 Best J. Slow death of the bleep: why hospital pagers won't die. *BMJ* 2021;372:684.

3 Menon R, Rivett C. Time-motion analysis examining of the impact of medic bleep, an instant messaging platform, versus the traditional pager: a prospective pilot study. *Digit Health* 2019;5:2055207619831812.

4 Byrd TF, Fancher KG, Liebovitz DM, *et al*. Trends in secure mobile communication technology use among hospitalists in north america, 2016–2021. *Health Policy Technol* 2022;11:100689.

5 Byrd TF, Speigel PS, Cameron KA, *et al*. Barriers to adoption of a secure text messaging system: a qualitative study of practicing clinicians. *J Gen Intern Med* 2022.

BMJ Health & Care Informatics

Novel machine learning model for predicting multiple unplanned hospitalisations

Paul Conilione ,[1] Rebecca Jessup ,[2,3] Anthony Gust [1]

To cite: Conilione P, Jessup R, Gust A. Novel machine learning model for predicting multiple unplanned hospitalisations. *BMJ Health Care Inform* 2023;**30**:e100682. doi:10.1136/bmjhci-2022-100682

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/bmjhci-2022-100682).

Received 07 October 2022
Accepted 06 March 2023

[1]Digital Health, Northern Hospital, Epping, Victoria, Australia
[2]Staying Well, Northern Hospital, Epping, Victoria, Australia
[3]School of Allied Health, Human Service and Sport, La Trobe University, Melbourne, Victoria, Australia

Correspondence to
Dr Paul Conilione;
paul.conilione@nh.org.au

ABSTRACT

Background In the Australian public healthcare system, hospitals are funded based on the number of inpatient discharges and types of conditions treated (casemix). Demand for services is increasing faster than public funding and there is a need to identify and support patients that have high service usage. In 2016, the Victorian Department of Health and Human Services developed an algorithm to predict multiple unplanned admissions as part of a programme, Health Links Chronic Care (HLCC), that provided capitation funding instead of activity based funding to support patients with high admissions.

Objectives The aim of this study was to determine whether an algorithm with higher performance than previously used algorithms could be developed to identify patients at high risk of three or more unplanned hospital admissions 12 months from discharge.

Methods The HLCC and Hospital Unplanned Readmission Tool (HURT) models were evaluated using 34 801 unplanned inpatient episodes (27 216 patients) from 2017 to 2018 with an 8.3% prevalence of 3 or more unplanned admissions in the following year of discharge.

Results HURT had a higher AUROC (84%, 95% CI 83.4% to 84.9% vs 71%, 95% CI 69.4% to 71.8%) than HLCC, that was statistically significant using Delong test at $p<0.05$.

Discussion We found features that appear to be strong predictors of admission risk that have not been previously used in models, including socioeconomic status and social support.

Conclusion The high AUROC, moderate sensitivity and high specificity for the HURT algorithm suggests it is a very good predictor of future multi-admission risk and that it can be used to provide targeted support for at-risk individual.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Case-finding algorithms for identifying patients at risk of unplanned readmissions traditionally have focused on detecting one or more admissions over a 30-day to 365-day period from discharge. This study focuses on patients who have more frequent admissions, and aims to predict three or more unplanned admissions within 365 days of an index admission. This allows for better targeting of patients that would otherwise use more resources. Accurately predicting those who represent the highest hospital use is likely to lead to greater healthcare cost savings.

WHAT THIS STUDY ADDS

⇒ This study presents an algorithm for identifying patients at risk of three or more unplanned admissions using not only clinical information, socioeconomic indicators and living arrangements, but also a novel cascading chronic condition feature.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ This study demonstrates the importance of using a mixture of clinical, demographic and activity-based features for predicting patient outcomes. Our algorithm outperformed a similar algorithm that used existing weighted scoring approaches. The study demonstrates that machine learning-based methods for identifying patients who would benefit from targeted intervention have great potential in improving health system sustainability.

BACKGROUND

Potentially preventable admissions include any hospitalisations for acute and chronic conditions that may have been avoided with earlier intervention and rehospitalisation within 30 days of discharge due to inadequate discharge and/or follow-up.[1] In many high-income countries, potentially preventable hospitalisations have become an indicator of health system performance.[2] Reported rates of preventable hospitalisation range from 5% to 79%.[3] This wide range reflects not only differences in the definition of preventable admissions, but also geographical and socioeconomic differences in population composition.[4] For hospitals, potentially preventable admissions increase hospital demand, lead to bed blocking and patient flow issues, and in Australia account for 10% of all occupied beds and more than 748 000 admissions per year.[5]

The cost of providing healthcare in most high-income countries is considered to be unsustainable and will likely be unaffordable by 2050 in the absence of major reforms.[6] Identifying and averting these preventable hospitalisations is important for not only improving individual health outcomes but in controlling burgeoning healthcare expenditure. Case finding algorithms to identify those at highest risk of preventable hospitalisations are emerging as a key initiative that may allow for targeted care to prevent deterioration and future admissions.[7] A highly sensitive and specific case-finding algorithm should be able to identify only those patients most likely to have high future healthcare costs or hospital resource use.

Internationally, there is a growing body of literature on algorithms that aim to predict the likelihood of future admissions using different models, including the traditional logistic regression model and survival analysis and more recently popular modelling using machine learning techniques.[8] Many approaches focus on patients at risk in specific disease categories such as chronic obstructive pulmonary disease (COPD),[9] stroke/Transient Ischaemic Attack (TIA),[10] diabetes[11] or heart failure.[12] Others focus on unplanned readmissions, usually within 30 days of discharge, for any-cause using non-linear models,[13] gradient boosted decision trees[14] or artificial neural networks.[15]

In 2016, the Victorian Department of Health and Human Services (DHHS) initiated the HealthLinks Chronic Care programme (HLCC) that provided an alternative capitated funding model for patients with chronic and complex health conditions who were at high risk of multiple unplanned admissions. A key component of the programme was the use of a predictive algorithm called the HLCC model. The HLCC model uses an index unplanned admission as a triggering event and then combines diagnostic information from that admission and demographic information to create a ‘risk score’ for the probability of another three or more admissions in the next 12 months. Patients who score above a threshold value determined by logistic analysis of historical data are eligible to be included in the HLCC programme, receiving targeted preventative care.[16] The HLCC risk score was found to have a sensitivity of 41% and specificity of 78% over the 2-year evaluation across five participating Victorian health service providers.[17] The low sensitivity suggests that there are potentially many patients who would benefit from targeted intervention who are not being identified by this algorithm and the moderate specificity suggests that efforts with targeted intervention was wasted on some individuals who would not have gone on to have a preventable admission. This paper describes the development, and content, of a machine learning case-finding prediction tool with a higher sensitivity and specificity for identifying patients that are at high risk of all-cause potentially avoidable admissions within 12 months of discharge in an Australian setting.

METHODS

Setting

This was a single-centred study based at Northern Health (NH). NH is the major provider of acute (410 beds), subacute (251 beds) and ambulatory specialist services in Melbourne’s north. Residents originate from more than 184 countries, speak more than 106 languages and have lower levels of income, educational attainment and health literacy, and higher rates of unemployment than state averages.[18] The emergency department at NH is the busiest in the state with over 100,000 presentations per year.[19]

Study design and data sources

Participants

Eleven years of historical NH acute Inpatient (IP) emergency admitted episodic level data was used from 1 July 2008 to 30 June 2019 to build and test a new model that we named the HURT (Hospital Unplanned Readmission Tool) model. Outpatient (OP) and emergency department (ED) data were also linked to the unplanned IP activity. In addition, the Index of Relative Socio-economic Advantage and Disadvantage (IRSEAD) from the Australian Bureau of Statistics (ABS) Socio-Economic Indexes for Areas (SEIFA) data set were linked to patient’s residential postal address.

An unplanned admission is an unexpected or sudden health issue or event that results in an emergency admission. We only included acute IP episodes where the patient was 18 years or older at admission, the admission was not related to mental health, obstetrics, oncology or renal dialysis and the patient did not die during the episode. If there were any records that contains missing values then they were discarded.

Table 1 presents the summary statistics of the data used including demographic information and the features used for the final model. Where the percentages are the proportion of separations with the given flag. The features are defined in table 2 later in the paper. These data were included as DHHS and other jurisdictions have these data readily available. Data such as pathology and pharmacy were not included as the DHHS does not have this.

Patient and public involvement

Patients and the public were not involved in the design of this work.

Variable selection for the HURT

Variable (feature) selection is a manual or automatic process by which variables that have the highest impact on model performance (in this case prediction of future unplanned admissions) are selected and variables that do not help learning are discarded.

The Boruta R package was used to develop the HURT. Boruta R uses a novel feature selection algorithm that finds all relevant variables, where relevant means variables that are found to be associated with unplanned emergency admissions. Boruta can use a range of decision trees to derive the importance of each feature. Extreme Gradient Boosting (XGBoost) was used to measure the feature importance in the Boruta algorithm with 200 maximum runs to ensure feature importance was fully resolved.

We also created a novel feature which we called a cascading chronic condition flag. If a patient was coded

Table 1 Summary statistics of the 11 years of data with mean value or percentage of separations with the relevant flag

Property	All	Readmitted <3	Readmitted ≥3
No of separations	206 714	192 679	14 035
No of patients	125 743	125 258	4730
Female %	50.3%	50.5%	47.8%
Average admission age	55.9	55	67.7
Admission age 61–90 flag	41.5%	39.7%	66.7%
Admission age 90+ flag	2.7%	2.6%	3.8%
Chronic condition (cascading) COPD flag	4.5%	3.6%	17.4%
Chronic condition (cascading) disorder due to tobacco flag	6.4%	5.2%	22.7%
Chronic condition (cascading) heart failure flag	5.4%	4.3%	19.5%
Complexity ≥3 flag	27.4%	25.7%	50.1%
Failed to attend ratio OP 365 days	8.7%	7.7%	21.5%
Marital status flag	22.7%	21.5%	39.2%
No OP bookings in past 365 days flag	65.5%	68.1%	30.3%
No of HIPs (any group)	0.22	0.17	0.83
No of non-admitted ED presentations	0.5	0.4	1.5
No of OP attended past year	1.5	1.3	4
Potential avoidable emergency admission flag	25.5%	24.3%	42.4%
IRSEAD decile within Australia	4.6	4.6	4
Total LOS unplanned episodes 180 days	4.6	4.1	10.7
Total LOS unplanned episodes 365 days	5.4	4.7	14.9
Total no unplanned eisodes past 365 days	1.6	1.5	3.6
Usual accommodation agecare, alone flag	2.6%	2.4%	6.2%

COPD, chronic obstructive pulmonary disease; ED, emergency department; HIP, Health Independence Program; IRSEAD, Relative Socio-economic Advantage and Disadvantage; LOS, Length of Stay; OP, emergency department.

with a chronic condition, all subsequent episodes were also flagged for this chronic condition (ie, if a patient was diagnosed with COPD, all following episodes would be flagged with this condition, when previously this would not occur if their admission had been for a different reason). Hence, this information can be used by the machine learning (ML) modelling to help predict future unplanned admissions.

Weighting variable importance

Over the past decade, 'black box' machine learning algorithms have been increasingly used in critical decision-making processes. However, because it is unclear or unknown how the machine learning algorithm decides there have been reports of adverse results in some fields.[20]

To overcome this problem, we used interpretable models that allow for an understanding of why the machine learning algorithm makes particular decisions on individual cases. The SHAP (Shapley Additive exPlanations) score was used as it assigns each variable an importance value for each decision outcome. The SHAP score can then be visualised to illustrate how the decision tree-based machine learning is making a given decision in an interpretable manner.[21]

Training and optimisation of the model

HURT is trained and tested on historical data where we know in advance if a patient had three or more unplanned admissions 1 year from the index admission. We define this as the 'target' for the model to be trained and tested on.

The XGBoost machine learning algorithm uses an ensemble (collection) of weak decision trees that are sequentially created to progressively improve (ie, boosting) the learning performance.[22] This has the advantage of quick training and has been shown to perform well on unseen test data. It also has the advantage of being able to handle unbalanced data where there are fewer positive cases (ie, patient returned three or more times in the future) compared with the negative cases (ie, patient did not return three or more times).

Like most machine learning algorithms, XGBoost has a set of training parameters that impact the final model performance. The parameter values that maximise performance cannot be determined by analysing the data only. These can only be found by trying different training parameters and measuring the model performance.[23] Hence, we performed hyperparameter optimisation by using a simple grid-search over a range of parameter

Table 2 List of variables in final model for predicting three or more unplanned admissions

Feature	Data type	Description
Admission age 61–90 flag	Binary	1 if Admitted age 61–90, 0 otherwise
Admission age 90+ flag	Binary	1 if admitted age ≥90, 0 otherwise
Chronic condition (cascading) COPD flag	Binary	Cascading chronic condition flag for COPD
Chronic condition (cascading) heart failure flag	Binary	Cascading chronic condition flag for heart failure
Chronic condition (cascading) disorder due to tobacco flag	Binary	Cascading chronic condition flag for disorder due to tobacco use
Complexity ≥3 flag	Binary	1 if number of body systems treated is 3 or more,[28] 0 otherwise.
Marital status flag	Binary	1 if divorced, widowed or separated, 0 otherwise
Total unplanned episodes past 365 days	Integer	In past 365 days prior to discharge
No of non-admitted ED presentations 365 days	Integer	Calculated for past 365 days from discharge
No of HIPs (any group)	Integer	Number of enrolments in any HIP group for past 180 days
No of OP attended	Integer	In past 365 days from IP discharge
Failed to attend ratio OP 365 days	Float	Calculated for past 180 and 365 days from discharge
No OP bookings flag past 365 days flag	Binary	1 if no OP bookings in past year, 0 otherwise
Usual accommodation age care, alone flag	Binary	1 if patient is living in age care facility or living alone, 0 otherwise
Potential avoidable emergency admission flag	Binary	1 if patient had ICD10 diagnosis code from,[29] 0 otherwise
IRSEAD Decile Within Australia	Integer	ABS Index of Relative Socio-economic Advantage and Disadvantage by postcode (0-low, 10-high)
Total LOS unplanned episodes 180 days	Integer	Calculated for past 180 days from discharge
Total LOS unplanned episodes 365 days	Integer	Calculated for past 365 days from discharge

ABS, Australian Bureau of Statistics; COPD, chronic obstructive pulmonary disease; ED, emergency department; HIP, Health Independence Program; ICD10, International Statistical Classification of Diseases and Related Health Problems, 10th Version; IP, inpatient; IRSEAD, Index of Relative Socio-economic Advantage and Disadvantage; LOS, Length of Stay; OP, outpatient.

values. The optimal XGBoost parameters where Eta=0.05, Max Depth=4, Gamma=0, Colsample_bytree=1, min_child_weight=2, Subsample=0.5, Nrounds=400.

The 11 years of historical data were divided into training and testing phases. The optimal model parameter values that produced the highest area under the reciever operator curve (AUROC) performance using 10-fold cross-validation on the training data (9 years, 171 913 separations, 98 527 patients) were used. Testing was performed on the episodes that were discharged in 2017–18 (1 year, 34 801 separations, 27 216 patients), but 2018–2019 data were needed to count the unplanned admissions up to 1 year from 2017 to 18 discharge. The training and testing phase were performed by using the Caret R-package.[24]

Final variables selected for the HURT

Table 2 provides an overview of the final 18 features selected for HURT from an initial set of 199 features. The definition of all features tested are available as online supplemental material 1.

The performance of the HURT was assessed retrospectively by calculating the AUROC, sensitivity which is the percentage of separations where the patient was correctly predicted to have three or more potentially avoidable admissions in the 12 months following their discharge. We also assessed the specificity of the model (the percentage of patient separations incorrectly predicted to have three or more unplanned admissions). The higher the sensitivity of the model, the more patients correctly identified and the less that will be 'missed' and have a potentially preventable readmission.

Comparison

The primary comparison of our algorithm is with the DHHS HLCC algorithm using AUROC, sensitivity and specifity. We also compare the decisions made by HURT and HLCC on the same separations and illustrate the differences by a Venn diagram. Of particular interest is the false-positives (FP) where a patient is falsely flagged as returning three or more times and will be offered support services. This means resources may be used for patients that were not going to return. The false-negative (FN) cases are of concern as these patients are not flagged, and will not be offered extra services, thus returning three or more times since discharge. This places a strain on hospital resources that could have been reduced but more importantly potential missing patients that may have deteriorated.

Given the lack of existing research using three or more unplanned admissions within 1 year of discharge. We also applied the previously described methodology

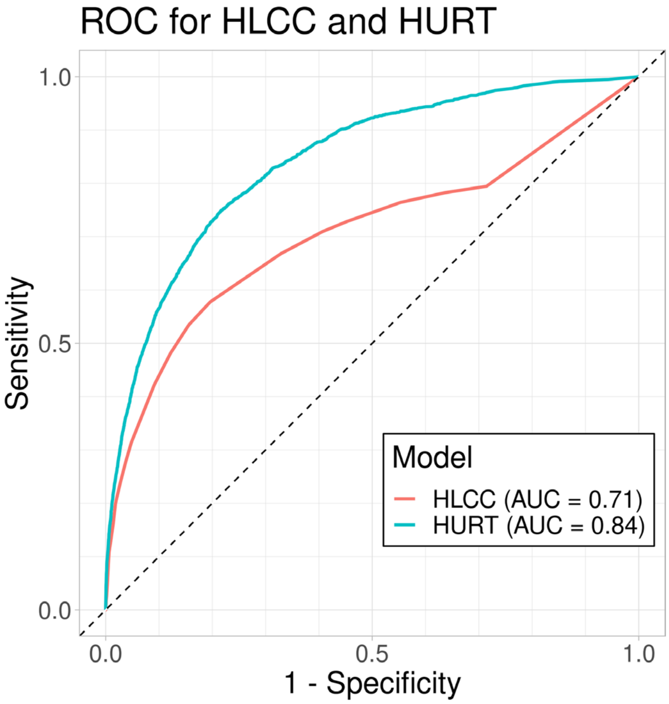


Figure 1 ROC test performance of HLCC and HURT models predicting three or more unplanned admissions. HLCC, HealthLinks Chronic Care; HURT, Hospital Unplanned Readmission Tool; ROC, receiver operating characteristic.

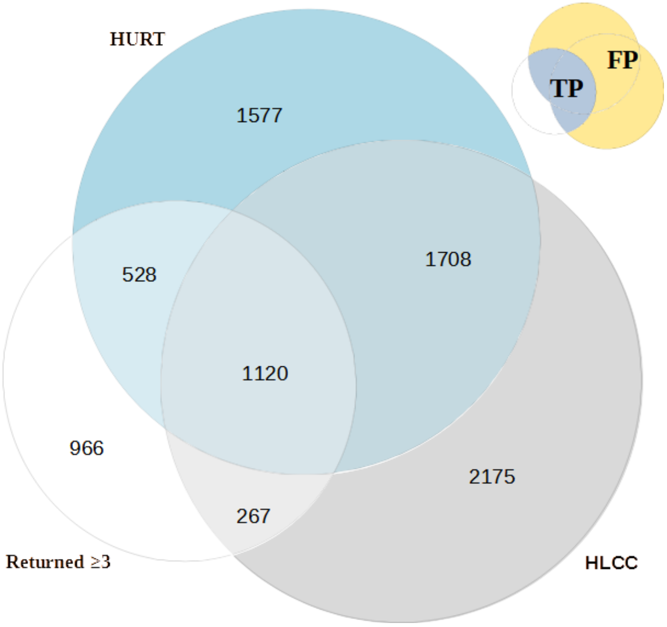


Figure 2 Venn diagram of number of separations that were predicted to return by machine learning, HLCC and overlap with the number of separations that actually returned three or more times. FP, false-positive; HLCC, HealthLinks Chronic Care; HURT, Hospital Unplanned Readmission Tool; TP, true-positive.

to predicting if a patient has one or more unplanned admissions within 1 year. Our model is compared with other case-finding algorithms that predict one or more unplanned admission over a year using weighted scores[25 26] and machine learning.[27] Even though this was not the focus of the research, it provided some reassurance of the methodology.

RESULTS

The optimised HURT model had a final test AUROC of 84% (95% CI 83.4% to 84.9%), while HLCC had an AUROC of 71% (95% CI 69.4% to 71.8%) (figure 1). The difference between HURT and HLCC ROC was statistically significant with Z=−22.6, p<0.001 (Delong test).

Using the confusion matrix in table 3, the HURT algorithm had a sensitivity of 57%, while HLCC had a sensitivity of 48%. The 9% difference was statistically significant with χ^2=85.03, p<0.001 (McNemar test). The HURT algorithm achieved 90% specificity, while HLCC had a specificity of 88%. The 2% difference was statistically significant with χ^2=94.99, p<0.001 (McNemar test).

The Venn diagram in figure 2 provides an overview of the number of hospital unplanned admissions that were predicted by each of the HLCC and HURT models in terms of true-positive (TP) and FP cases. The number of separations that were predicted correctly (ie, the overlap between 'returned≥3', HURT and HLCC) are TP cases (HURT: 528, both: 1120, HLCC: 267). Where HURT has 261 more separations correctly classified compared with HLCC. While the FP cases (HURT: 1577, both: 1708, HLCC: 2175) show the HURT has 598 fewer FPs compared with HLCC. Both models missed 966 positive cases.

The 18 most important variables for predicting admission can be grouped into three: demographics (particularly age and marital status), medical conditions (complexity and cascading chronic conditions, in particular COPD and chronic cardiac failure) and

Table 3 Comparison of the HURT and HLCC algorithms in identifying patients at risk of three or more unplanned readmissions

Algorithm	HLCC		HURT	
Prediction	Readmitted ≥3	Readmitted <3	Readmitted ≥3	Readmitted <3
Readmitted ≥3	1387 (TP)	1494 (FP)	1648 (TP)	1233 (FP)
Readmitted <3	3883 (FN)	28 037 (TN)	3285 (FN)	28 635 (TN)
Result	Sensitivity 48%	Specificity 88%	Sensitivity 57%	Specificity 90%

FN, false-negative; FP, false-positive; HLCC, HealthLinks Chronic Care; HURT, Hospital Unplanned Readmission Tool; TN, true-negative; TP, true-positive.

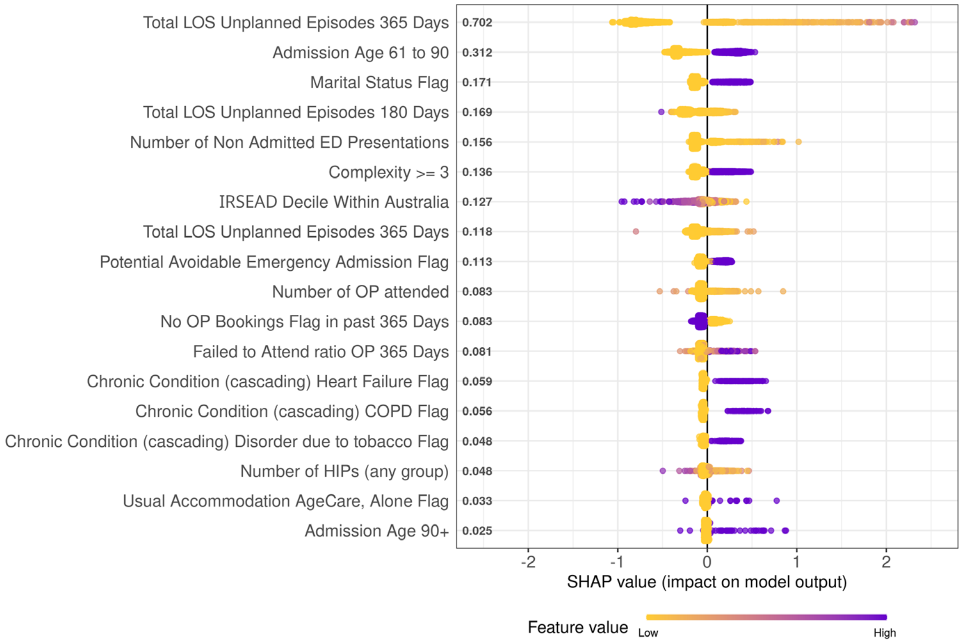


Figure 3 SHAP plot of impact of each feature on decision of XGBoost model. COPD, chronic obstructive pulmonary disease; ED, Extreme Gradient Boosting; HIP, Health Independence Program; LOS, Length of Stay; OP, outpatient; IRSEAD, Index of Relative Socio-economic Advantage and Disadvantage; SHAP, Shapley Additive exPlanations; XGBoost, Extreme Gradient Boosting.

past resource use (unplanned admissions, avoidable emergency presentations and failure to attend OP appointments). Figure 3 provides a SHAP plot of each of the 18 variables.

Tables 4 and 5 present the test AUROC, sensitivity and specificity of the proposed algorithm and other models both in Australian and internationally for comparison along with the 95% CIs. Not all the referenced papers provide full details on the data sizes and performance values for their models.

DISCUSSION

The HURT algorithm had an AUROC of 84%, sensitivity of 57% and specificity of 90%. In the model, the 2% higher specificity for the HURT over the HLCC translated into 598 fewer FP and 261 more TP predictions in the 12-month time frame. The HURT algorithm also flagged more patients that would have benefitted from targeted services who went on to have two or less unplanned admissions over 12 months.

Even though these findings are for a tertiary hospital in the state of Victoria, there are still lessons that can be applied to the broader healthcare system across Australia and internationally. In the local Australian context, the Independent Health and Aged Care Pricing Authority may apply penalties for hospitals that treat what are deemed avoidable readmissions (less than 30 days). As the HURT model has a higher specificity than other Australian models, it may be a more cost-effective tool for Australian hospitals to use as it will select less FP, and therefore, prevent hospitals who use this model from being avoidably penalised.

Researchers based in the UK have developed a number of case-finding algorithms[25–27] over the years. Direct

Table 4 Summary of test performance for predicting three or more unplanned admissions within 1 year of discharge for different case finding algorithms

						Target: 3 or more unplanned in 12 months		
Method	Country	Model	Data	Total separations (patients)	Test separations (patients)	AUROC (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)
HURT	Australia	XGBoost	IP, ED, OP, ABS	206 714 (125 743)	34 801 (27 216)	84.2% (83.4 to 84.9)	57.2% (55.4 to 59.0)	89.4% (89.4 to 90.0)
HLCC @ NH	Australia	Weighted score	IP, ED	206 714 (125 743)	34 801 (27 216)	70.5% (69.4 to 71.8)	48.1% (46.3 to 50.0)	87.8% (87.5 to 88.2)
HLCC Victoria[17]	Australia	Weighted score	IP, ED	N/A (N/A)	N/A (N/A)	N/A (N/A)	41%* (N/A)	N/A (N/A)

*Recall was 78%.
ABS, Australian Bureau of Statistics; AUROC, Area Under the Reciever Operator Curve; ED, emergency department; HLCC, HealthLinks Chronic Care; HURT, Hospital Unplanned Readmission Tool; IP, inpatient; N/A, not available; OP, outpatient.

Table 5 Summary of test performance predicting one or more unplanned admissions within 1 year of discharge for different case finding algorithms

Method	Country	Model	Data sources	Total separations (patients)	Test separations (patients)	Target: 1 or more unplanned in 12 months		
						Auroc (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)
HURT	Australia	XGBoost	IP,ED, OP, ABS	206 714 (125 743)	34 801 (27 216)	74.9% (74.2 to 75.5)	39.2% (38.2 to 40.2)	90% (89.5 to 90.4)
Billings *et al*[25] 2013	UK	Weighted score	IP, ED, OP, GP	N/A (1 836 099)	N/A (N/A)	78%* (N/A)	42%† (N/A)	N/A (N/A)
QAdmission Score[26]	UK	Weighted score	IP, ED, OP, GP	12 957 648 (4 870 488)	N/A (N/A)	Women: 77.3% (77.1 to 77.8) Men: 77.6% (77.4 to 77.8)	39%	N/A
SPARRA V4[27]	UK	Ensemble (ANN, RF, XGB, GLM, NB)	IP, ED, OP, GP	12 957 648 (4 870 488)	4 300 000 (N/A)	80.1% (SE: 0.023)	~52% ‡ (N/A)	~90%‡ (N/A)

*Patient-level performance.
†Recall was 78%.
‡Estimated from figure 2 (a).[27]
ABS, Australian Bureau of Statistics; ANN, Artificial Neural Network; ED, emergency department; GLM, Generalised Linear Model; GP, general practitioner; HURT, Hospital Unplanned Readmission Tool; IP, inpatient; N/A, not available; NB, Naive Bayes; OP, outpatient; RF, Random Forest; XGBoost, Extreme Gradient Boosting.

comparisons with some of these other models is difficult given data scientists use different datasets (both in terms of data captured and patient cohorts), and different definitions of an unplanned admission and benchmarks for what is considered an acceptable number of these within a 12-month period. The UK models use one or more unplanned admissions of any cause as their benchmark,[10 25–27] with the SPARRA V4 demonstrating the highest AUROC (80%) with a sensitivity of approximately 52% and a specificity of approximately 90%.[27] Where sensitivity and specificity were estimated from figure 2 (a) ROC plot.[27] Future algorithm research would benefit from application of consistent definitions so that developed algorithms may be tested and applied within different healthcare contexts (rural, remote and metropolitan) and countries.

Of particular interest in this study were the results from the SHAP scores for the importance of each feature in the HURT algorithm. Higher numbers of unplanned hospital admissions and ED admissions in the past year are shown to be important predictive features of future unplanned readmissions. In addition, lower socioeconomic status and lack of social support was predictive of unplanned readmissions, which was in agreement with SPARRA who used the Scottish Index of Multiple Deprivation using SHAP scores.[21] Both QAdmission[26] and SPARRA[27] found pathology and medication history to be an important feature for prediction of readmission, which would explain their higher performance. These data were deliberately left out so that other jurisdictions could build our model. Our next version will include this data.

The limitation of this study is that it focuses on the application of machine learning to the problem of predicting if a patient would have unplanned readmissions given current and historical information for an index admission. Hence, we only examined the performance of HURT on NH data and compared to the HLCC which was used in several Victorian health services. The model has not been subject to external validation and may not work well in non-tertiary (hospital) sites. Further work will involve multiple phases. The first phase will be to evaluate HURT within a live production system, both in terms of classification performance (eg, sensitivity) and operationally (cost savings, cohort selection). The second phase will draw on the first to improve the HURT as it is a part of a broader system that will be evaluated and optimised. Patient cohorts will be examined for FP/FN to determine any striking features that can be used or enhanced to improve ML identification of patients that will have unplanned admissions. Finally, the aim is to use general practice, pharmacy and pathology data, patient survey data and sensor in the home to better predict patients likely to readmit. These data were not included in the current approach because it is not available to the Victoria Department of Health.

CONCLUSIONS

We developed the HURT based on the XGBoost ML algorithm. We also created novel features from hospital medical and administrative data called Cascading Chronic Conditions. The HURT algorithm was compared to the Victorian Department of Health HLCC scoring method for identifying patients at risk. The HURT model was found to have AUROC of 84%, sensitivity of 57% and specificity with 90%, 14%, 9% and 2% better than the HLCC, respectively. Future research will use pathology and pharmacy data with the aim of improving model performance.

Twitter Rebecca Jessup @Jessuprl

Acknowledgements We would like to thank Professor Don Campbell and Keith Stockman from the Staying Well program at Northern Health for their support and feedback.

Contributors AG, conceived of the study. AG conducted the thematic analysis, PC conducted the machine learning. PC wrote the first draft of the manuscript. All authors provided critical feedback related to the study results and written manuscript and approved the final manuscript for publication.

Funding This study was supported by the Better Care Innovation Fund, Safer Care Victoria Number: IF19112.

Disclaimer The funding body had no role in the design of the study, collection, analysis and interpretation of the data and in writing the manuscript.

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval Approval was granted by the Northern Health Low Risk Ethics Committee. Local reference number: ALR 53.2020-69129.

Provenance and peer review Not commissioned; internally peer reviewed.

Data availability statement Data are available on reasonable request. The datasets generated and/or analysed during the current study are not publicly available due to the confidential nature of the patient data. However, researchers with inquiries related to the study methods or results should contact the corresponding author at paul.conilione@nh.org.au.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

ORCID iDs
Paul Conilione http://orcid.org/0000-0001-9913-8824
Rebecca Jessup http://orcid.org/0000-0003-2211-5231
Anthony Gust http://orcid.org/0000-0003-0509-2461

REFERENCES

1 Katterl R, Anikeeva O, Butler C, *et al*. *Potentially avoidable hospitalisations in australia: causes for hospitalisations and primary health care interventions*. 2012.
2 Spencer J. *National health and hospitals network agreement*. Australia: Council of Australian Governments, 2011.
3 van Walraven C, Bennett C, Jennings A, *et al*. Proportion of hospital readmissions deemed avoidable: a systematic review. *CMAJ* 2011;183:E391–402.
4 Falster MO, Jorm LR. A guide to the potentially preventable hospitalisations indicator in Australia. In: *Australian Commission on Safety and Quality in Health Care*. Sydney: Centre for Big Data Research in Health, University of New South Wales in consultation with Australian Commission on Safety and Quality in Health Care and Australian Institute of Health and Welfare, 2017: 34.
5 AIHW. Potentially preventable hospitalisations in australia by age groups and small geographic areas, 2017–18 [Overview]. 2019. Available: https://www.aihw.gov.au/reports/primary-health-care/potentially-preventable-hospitalisations/contents/about
6 OECD. *Fiscal sustainability of health systems: bridging health and finance perspectives*. 2015: 264.
7 Pang RK, Srikanth V, Snowdon DA, *et al*. Targeted care navigation to reduce hospital readmissions in “at-risk” patients. *Intern Med J* 29, 2021.
8 Artetxe A, Beristain A, Graña M. Predictive models for hospital readmission risk: a systematic review of methods. *Comput Methods Programs Biomed* 2018;164:49–64.
9 Min X, Yu B, Wang F. Predictive modeling of the hospital readmission risk from patients’ claims data using machine learning: a case study on COPD. *Sci Rep* 2019;9:2362.
10 Hung L-C, Sung S-F, Hu Y-H. A machine learning approach to predicting readmission or mortality in patients hospitalized for stroke or transient ischemic attack. *Applied Sciences* 2020;10:6337.
11 Hammoudeh A, Al-Naymat G, Ghannam I, *et al*. Predicting Hospital readmission among diabetics using deep learning. *Procedia Computer Science* 2018;141:484–9.
12 Ashfaq A, Sant’Anna A, Lingman M, *et al*. Readmission prediction using deep learning on electronic health records. *J Biomed Inform* 2019;97:103256.
13 Yang C, Delcher C, Shenkman E, *et al*. Predicting 30-day all-cause readmissions from hospital inpatient discharge data. 2016 IEEE 18th International Conference on e-Health Networking, Applications and Services (Healthcom); Munich, Germany: IEEE, Munich, Germany.
14 Maali Y, Perez-Concha O, Coiera E, *et al*. Predicting 7-day, 30-day and 60-day all-cause unplanned readmission: a case study of a Sydney Hospital. *BMC Med Inform Decis Mak* 2018;18:1.
15 Jamei M, Nisnevich A, Wetchler E, *et al*. Predicting all-cause risk of 30-day Hospital readmission using artificial neural networks. *PLOS ONE* 2017;12:e0181173.
16 Division HaW. HealthLinks chronic care evaluation. In: *Summary of implementation and outcomes for 2016–17*. 1 Treasury Place, Melbourne: Victorian Government, 2019: 45.
17 Good N, Niven P, Li J, *et al*. HealthLinks: chronic care evaluation: summary report. 1 treasury place. Melbourne, Australia Victorian Government; 2022. 110.
18 Australian Bureau of Statistics. Socio-economic indexes for areas (SEIFA) 2016: commonweath government. 2016. Available: https://www.abs.gov.au/ausstats/abs@.nsf/mf/2033.0.55.001
19 Jessup RL, Bramston C, Beauchamp A, *et al*. Impact of COVID-19 on emergency department attendance in an australia hospital: a parallel convergent mixed methods study. *BMJ Open* 2021;11:e049222.
20 Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell* 2019;1:206–15.
21 Lundberg SM, Lee S-I. *A unified approach to interpreting model predictions*. Red Hook, NY, USA: Curran Associates Inc, 2017.
22 Chen T, Guestrin C. XGBoost: A scalable tree boosting system. *KDD ’16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 2016.
23 Chicco D. Ten quick tips for machine learning in computational biology. *BioData Min* 2017;10:35.
24 Kuhn M, Wing J, Weston S, *et al*. CARET: classification and regression training [internet]. 2020. Available: https://CRAN.R-project.org/package=caret
25 Billings J, Georghiou T, Blunt I, *et al*. Choosing a model to predict hospital admission: an observational study of new variants of predictive models for case finding. *BMJ Open* 2013;3:e003352.
26 Hippisley-Cox J, Coupland C. Predicting risk of emergency admission to hospital using primary care data: derivation and validation of QAdmissions score. *BMJ Open* 2013;3:e003482.
27 Liley J, Bohner G, Emerson SR, *et al*. Development and assessment of a machine learning tool for predicting emergency admission in scotland. *Public and Global Health* [Preprint].
28 Hart M, Dean D. *How complex is that patient*. The Health Roundtable Limited, 2006.
29 Martin C, Stockman K, Hinkley N, *et al*. Multimorbidity and acute potentially preventable diagnoses in healthlinks chronic care (HLCC) dandenong cohort. A work in evolution. *Int J Integr Care* 2021;20:173.

BMJ Health & Care Informatics

Analysis of 'One in a Million' primary care consultation conversations using natural language processing

Yvette Pyne ,[1] Yik Ming Wong,[2] Haishuo Fang,[2] Edwin Simpson[2]

To cite: Pyne Y, Wong YM, Fang H, *et al*. Analysis of 'One in a Million' primary care consultation conversations using natural language processing. *BMJ Health Care Inform* 2023;**30**:e100659. doi:10.1136/bmjhci-2022-100659

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/bmjhci-2022-100659).

Received 05 August 2022
Accepted 30 March 2023

[1]Bristol Medical School, University of Bristol Centre for Academic Primary Care, Bristol, UK
[2]Intelligent Systems Labs, University of Bristol, Bristol, UK

Correspondence to
Dr Yvette Pyne;
yvette@digitalgp.net

ABSTRACT

Background Modern patient electronic health records form a core part of primary care; they contain both clinical codes and free text entered by the clinician. Natural language processing (NLP) could be employed to generate these records through 'listening' to a consultation conversation.

Objectives This study develops and assesses several text classifiers for identifying clinical codes for primary care consultations based on the doctor–patient conversation. We evaluate the possibility of training classifiers using medical code descriptions, and the benefits of processing transcribed speech from patients as well as doctors. The study also highlights steps for improving future classifiers.

Methods Using verbatim transcripts of 239 primary care consultation conversations (the 'One in a Million' dataset) and novel additional datasets for distant supervision, we trained NLP classifiers (naïve Bayes, support vector machine, nearest centroid, a conventional BERT classifier and few-shot BERT approaches) to identify the International Classification of Primary Care-2 clinical codes associated with each consultation.

Results Of all models tested, a fine-tuned BERT classifier was the best performer. Distant supervision improved the model's performance (F1 score over 16 classes) from 0.45 with conventional supervision with 191 labelled transcripts to 0.51. Incorporating patients' speech in addition to clinician's speech increased the BERT classifier's performance from 0.45 to 0.55 F1 (p=0.01, paired bootstrap test).

Conclusions Our findings demonstrate that NLP classifiers can be trained to identify clinical area(s) being discussed in a primary care consultation from audio transcriptions; this could represent an important step towards a smart digital assistant in the consultation room.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Natural language processing (NLP) has the potential to revolutionise clinical specialties that rely on free text such as primary care which extensively uses electronic health records.

⇒ Existing NLP tools are focused on classifying free text created by health professionals or generating free text from predefined clinical data.

⇒ The creation of a tool to classify a clinical consultation based on the conversation that occurs in it could have a significant positive effect on clinician workload and could form part of the tools used in an 'augmented consultation'.

WHAT THIS STUDY ADDS

⇒ This study is the first to analyse and classify primary care consultations from the conversations that took place between doctors and patients.

⇒ This study develops and assesses the efficacy of several NLP classifiers, including recent pretrained deep neural networks, for classifying verbatim medical conversation transcripts, which use very different language to clinical notes, and for which extremely limited training data are available.

⇒ This study identifies limitations of the existing healthcare datasets and tools containing primary care free text and makes recommendations for further avenues of research and appropriate data sources.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ This study highlights the importance of building datasets of clinical conversations and other healthcare-based natural language sources for use in clinical research.

⇒ This study suggests several further research topics combining the fields of clinical primary care and machine learning.

INTRODUCTION

Technology is becoming increasingly pervasive in primary care[1] and a significant proportion of a clinician's day is spent interacting with the patient electronic health record (EHR). EHRs are a form of 'handover', either to another health professional, or to the same clinician when they meet the patient again; the records also provide key evidence in legal cases and are used for performance targets (such as the UK National Health Service Quality and Outcomes Framework) and billing (in the USA). EHRs incorporate free text and clinical codes such as SNOMED-CT, ICD (International Classification of Diseases) or Read codes. Historically, EHRs have been for clinicians only, but incoming UK legislation will open these records to be viewed by patients as well. For all these reasons, it is vital that clinical notes and their associated codes are accurate and complete.

Modern EHR systems used in UK primary care (such as EMIS, SystmOne and Vision) can also direct clinicians to clinically relevant local or national guidelines if the clinician enters an appropriate clinical code. However, clinical codes are often associated with the diagnosis rather than the presenting complaint so may only be entered at the conclusion of the consultation or even after the patient has left. Writing EHR notes or entering clinical codes during a consultation can be disruptive as the clinician has to focus on data capture rather than the patient.[2 3] Motivated by this, we investigated the first steps towards a natural language processing (NLP)[4] application that can 'listen' to a conversation between general practitioner (GP) and patient and automatically recommend clinical codes.

NLP has previously been applied to healthcare in a wide range of applications; for example, to process and analyse patient feedback,[5] identify risk factors,[6] symptoms and treatments,[7] or suspected disease[8] from clinical notes, or even to generate notes automatically from structured hospital data.[9] The technology to transcribe speech to text already exists in tools such as 'Otter.ai',[10] which could enable text processing of clinical conversations. However, the systematic evaluation of the use of NLP for interpreting conversations between clinicians and patients is lacking.[11–14]

We treated the task of assigning clinical codes to transcripts as text classification, which can be addressed using supervised learning. However, training data are in short supply, and recent NLP approaches based on deep learning are data hungry. This research assessed a series of text classifiers trained with small datasets to identify clinical codes associated with real-life GP–patient consultations. Our objectives were to evaluate: (1) the performance of different kinds of text classifiers; (2) the effect of training classifiers using existing medical code descriptions rather than example consultations; (3) the contribution of patients' speech to correct classifications in addition to the clinician's speech and (4) opportunities for improving the classifiers in future.

METHODS

Data sources

'One in a Million' dataset

The 'One in a Million' (OIAM) dataset[15] contains 300 video and audio recordings and verbatim transcripts of real clinical consultations conducted in 12 GP practices around Bristol in English with adult patients with permission in place for reuse. These consultations are associated with one or more International Classification of Primary Care (ICPC-2) clinical problem codes assigned by human coders. Both anonymised transcripts and ICPC-2 codes were available for 239 consultations.[16] A fictional but representative part of a consultation transcript is shown in online supplemental appendix A.

ICPC-2: ICPC-2 code descriptions

This is a primary care focused set of approximately 1300 low-level codes related to clinical problems that are grouped into 17 high level chapters or codes associated with clinical problem areas such as 'urinary' or 'circulatory'.[17] The ICPC-2e-V.7.0 comma separate values file[18] was used to create a data dictionary of high-level codes associated with relevant words for that group of conditions.

National Institute for Health and Care Clinical Knowledge Summaries

We created a National Institute for Health and Care Clinical Knowledge Summaries (NICE CKS) 'Health Topics' dataset using the 'Web Scraper.io' Google Chrome extension on 29 July 2021 from the publicly available web resource covering over 370 clinical topics.[19] For each health topic, we considered text from sections: 'Causes', 'Definition', 'Diagnosis', 'Clinical features', 'History', 'Presentation', 'Signs and symptoms' and 'When to suspect'. The clinical author mapped each NICE CKS topic to one or more ICPC-2 codes (see online supplemental appendix B: ICPC-2 codes and consultations). Then, for each ICPC-2 code, all the related CKS health topics were concatenated into a single document corresponding to that ICPC-2 code. While the ICPC-2 descriptions contain lists of relevant keywords, CKS health topics contain complete sentences that may convey additional information such as descriptions of symptoms.

Training the NLP classifiers

We initially used the OIAM dataset to train and test a series of classifiers using standard supervised learning (objective (1)). We held out a stratified sample of 20% (48 transcripts) of OIAM as a test set, using the remainder (191 transcripts) for training. Hyperparameter tuning was performed using fivefold cross-validation on the training split (see online supplemental appendix D).

Supervised learning requires a training dataset containing sufficiently representative examples for each class label, yet our training set contains only a small number of example consultations per code. We, therefore, introduced a second approach, 'distant supervision', that used the ICPC-2 code descriptions and NICE CKS datasets as training examples and tested the classifiers on the OIAM dataset (objective 2). We also tested excluding the 'A: General' classification as it includes a wide spectrum of clinical conditions from 'pain general/multiple sites' to 'viral disease other', and thus assigning the code was unlikely to aid GPs and may confuse the classifiers. Finally, we analysed distant supervision performance considering only the GP's half of the conversation to determine whether transcribing patient's speech is beneficial (objective 3).

To assess classifier performance, we used the macroaverage precision (equivalent to positive predicted value; the fraction of labels assigned by the classifier that were correct), recall (also called 'sensitivity'; the fraction of true labels predicted by the classifier) and F1 score (the harmonic mean of precision and recall).

As a baseline, we assigned labels at random, allowing multiple labels per transcript. We tested shallow, data-efficient classifiers: naïve Bayes (NB), as a linear classifier; support vector machine (SVM) with RBF kernel, a non-linear classifier that performs well with high dimensional feature vectors, such as those used to represent text (see below); and nearest centroid with Euclidean distance as a lightweight clustering-based classifier. While there are many other alternatives, the chosen methods represent broad types of classifier and allowed us to determine the suitability of classifiers with increasing complexity (part of objective (1)). The NB and SVM classifiers were run in 'multilabel' and 'multiclass' classification modes:

- Multilabel: for each possible ICPC-2 code, we train a binary classifier to assign either 'yes' or 'no' per consultation, so that more than one code can be assigned to the consultation.
- Multiclass: we train one classifier to assign the single most likely ICPC-2 code to the consultation. In training, we select the first code for each consultation, with codes sorted alphabetically.

In both modes, classifiers were evaluated on the complete test dataset using the same metrics. For consultations with more than one ICPC-2 code, the correct set of labels must be predicted to achieve perfect recall. As there are 110 consultations with more than one label, this puts a ceiling on the recall of the multiclass approach. However, the training data are more balanced, which may lead to better recall than the multilabel setup, where the training data for each binary classifier contains only a small minority of positive examples. Precision could also be higher as the multiclass mode directly compares classes that are easy to confuse.

For the shallow classifiers, we removed stopwords from the consultation transcripts before processing them. Considering the choice of stopwords as part of objective (1), we tested 3 sets: 318 'English' stopwords (from sklearn's default ENGLISH_STOP_WORDS); 203 'medical' stopwords[20] and 61 'custom' stopwords (see online supplemental appendix C: custom stopword dictionary). We encoded each transcript, ICPC-2 code description and CKS health topic as a feature vector containing the counts of the 5000 most frequent unigrams (individual words) and bigrams (consecutive pairs of words).

We also trialled recent deep learning classifiers that leverage a pretrained transformer, PubMedBERT,[21] a variant of BERT[22] that was pretrained on biomedical text (objective (1)). PubMedBERT encodes text into dense vector representations that take word order into account and include medical terms not present in our training examples. We tested a 'conventional BERT' classifier, in which we fine-tuned a classification head on top of PubMedBERT (multiclass mode). For distant supervision, we compared this to two BERT setups designed for training with very few examples: using next sentence prediction (NSP) to compare the text to a prompt containing the name of each class (multilabel mode); and using masked language modelling (MLM) to predict the category name by filling in the blank word in a prompt (multiclass)[23] ; both used 'this is a problem of ___' as a prompt. We hypothesised that the BERT approaches would outperform shallow classifiers thanks to their pretrained language representations, and that MLM would perform best as it reuses the pretraining task, so does not need to learn new classifier layers from scratch. Since BERT has a length limit of 512 tokens, transcripts and CKS topics were broken into multiple documents consisting of complete sentences. For training, all chunks were assigned the corresponding ICPC-2 training label. For prediction, we took the union of labels predicted for each of the chunks.

RESULTS

The consultation and patient demographics for the OIAM dataset are given in table 1, and the number of transcripts with multiple labels is shown in figure 1.

Objective (1): types of NLP classifiers

Table 2 shows the results for classifiers trained on OIAM transcript texts, with best performances highlighted in bold. As the held-out test set is small, we include the results of fivefold cross-validation over the larger training set. Nearest centroid is the best shallow classifier. Multiclass NB clearly outperforms SVM, while BERT provides substantial improvements all round. Compared with multilabel mode, multiclass classifiers have higher precision. However, recall and F1 are lower for multiclass SVM, while they are higher for multiclass NB, despite being unable to assign multiple codes to a single transcript. The baseline slightly outperforms multilabel NB on the test set and is competitive with some other shallow methods.

A comparison of F1 scores with different stopwords is shown in table 3, with the best choice for each classifier in bold, corresponding to the results in Table 2. Removing English or medical stopwords is helpful, while removing the words in all three stopword lists is most effective.

Objective (2): distant supervision

Table 4 compares F1 scores for different stopword lists with distant supervision. With CKS, the combined list is again most effective, but medical stopword removal is detrimental with ICPC-2 descriptions. Since ICPC-2 descriptions contain keywords rather than prose, any medical stopwords included by the authors of the descriptions may be part of informative key phrases that should not be removed.

Table 5 compares performance on the OIAM training set using distant supervision with the ICPC-2 code descriptions and NICE CKS topics. NB performs best with ICPC-2 supervision, in this case outperforming nearest centroid. BERT does not match the performance of NB multiclass on this small training set and conventional BERT fails to learn at all. BERT variants perform better with CKS than ICPC-2 as PubMedBERT was pretrained to process prose, rather than keywords. Combining both distant

Table 1 Details of the OIAM dataset used in this work, with patient information for the complete dataset

ICPC-2 code	No of transcripts	%
A: General	14	5.9
B: Blood, blood forming	8	3.3
D: Digestive	44	18.4
F: Eye	5	2.1
H: Ear	11	4.6
K: Circulatory	32	13.4
L: Musculoskeletal	65	27.2
N: Neurological	20	8.4
P: Psychological	50	20.9
R: Respiratory	37	15.5
S: Skin	32	13.4
T: Metabolic, endocrine, nutritional	24	10.0
U: Urinary	18	7.5
W: Pregnancy, family planning	11	4.6
X: Female genital	14	5.9
Y: Male genital	7	2.9
Total ICPC-2 code labels	392	164
Total unique consultations	239	100
No of ICPC-2 codes assigned to a consultation (see figure 1)		
0	2	1
1	128	53
2	62	26
3	40	17
4+	8	3
Duration (minutes)		
<5	13	5.4
5–10	79	33.1
10–15	82	34.3
15–20	52	21.8
20–35	13	5.4
Dataset statistics below are for the original patient sample of N=334.[16] This information was not available to compute for the N=239 subset in our experiments	No of patients	%
Sex		
Female	212	63.5
Male	122	36.5
Age		
18–34	91	27.2
35–54	94	28.1
55–74	99	29.6
≥75	36	10.8
Not reported	14	4.2
Ethnic group		
White	291	87.1
Other	43	12.9
IMD (Indices of Multiple Deprivation) quintile		
1st (least deprived)	106	31.7
2nd	54	16.2
3rd	35	10.5
4th	53	15.9
5th (most deprived)	84	25.1
Data unavailable	2	0.6

ICPC-2, International Classification of Primary Care-2; OIAM, One in a Million.

supervision sources does not improve performance for any of the methods (table 6).

Table 2 also shows that removing the option of assigning class A causes a collapse in performance with BERT NSP and MLM with ICPC-2 descriptions, and nearest centroid with either supervision source, while NB is improved slightly.

Table 2 shows test set performance with the most successful distant supervision source for each classifier. In comparison with standard supervision, the performance improves substantially for most classifiers, validating the use of external sources for distant supervision.

The NB model allows direct interpretation of the important features for classification. The wordclouds in figure 2 show the unigrams and bigrams for each class, weighted by the probability of the class given the feature, as learnt by NB (multiclass) from ICPC-2 descriptions. The informative features correspond well with medical terms in each category, but we do not see colloquial terms that may be used in conversation, or expressions longer than two tokens. Therefore, classifiers may benefit from augmenting ICPC-2 descriptions with alternative terms and phrases (objective 4, future improvements).

Objective (3): contribution of patient speech transcripts

Table 6 shows that using only the GP's part of the transcript reduces performance of most classifiers, indicating that patients provide useful information that is not contained in the GP's speech. The CIs only indicate strong evidence of a performance difference for BERT conventional, hence the finding may require investigation with a larger dataset.

DISCUSSION

We evaluated a range of text classifiers, achieving the highest F1 score on the test set of 0.51 for conventional

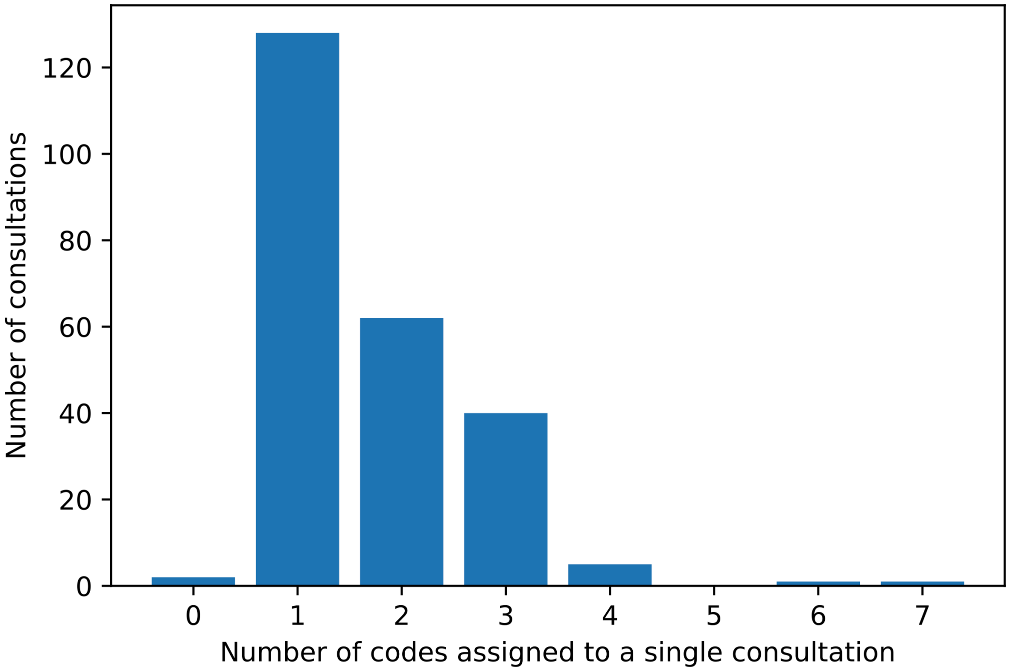


Figure 1 Distribution of consultations with multiple labels.

BERT, with recall at 56% and precision at 55%, substantially better than n-gram-based classifiers (objective 1). This classifier was trained on medical code descriptions, which outperformed standard supervision with a training set of 191 transcripts (those with no missing data such as codes, transcripts or notes) with F1=0.45 (objective 2). When patients' speech transcripts were excluded, the performance also dropped from F1=0.55 to 0.45 showing that is beneficial to capture the complete conversation (objective 3). Below, we identify specific ways to further improve the classifiers (objective 4).

More work is required to determine whether classifiers with this level of performance could usefully assist clinicians. Our scores are at the lower end of results for comparable multiclass text categorisation tasks,[24] which achieved between 53% and 86% average accuracy using a RoBERTa classifier with 100 training examples, and substantially lower than BERT for intent classification on dialogue benchmarks,[25] which achieves almost 93% accuracy with 10 training examples. Future work could, therefore, draw on these related tasks to identify improvements to the classifiers.

NB was competitive with BERT suggesting that unigrams and bigrams provide strong signals about health topics, and that datasets on the scale of OIAM may be insufficient to make full use of deep models. Against our expectations, conventional BERT was marginally the strongest, outperforming BERT MLM on the test set. The BERT models are costly to run (several hours GPU training for all BERT variants vs a few seconds with NB; testing takes in around 100 times longer), although this may not be an issue if training is performed only once before deploying the model. Future work could investigate replacing PubMedBERT with other domain-specific pretrained models (such as BioBERT[26] and ClinicalBERT[27]). Extremely large language models (LLMs) may also offer improved few-shot learning, although extensive prompt engineering is required and computational costs are huge. These LLMs could potentially generate explanations of their decisions that could bring relevant parts of the conversation to a doctor's attention.

The multilabel classifiers did less well than the multiclass classifiers, possibly because their training data was highly imbalanced (harming recall) or because multiple labels were assigned in cases where only one of the labels should have been chosen (hurting precision). However, given the complexity and breadth of primary care consultations, any effective classifier would need to be able to suggest multiple medical areas, so multilabel methods must be a focus for future research.

Given the low numbers of examples of some codes (eg, only five consultations were coded as 'F: eye'), overfitting was an issue for supervised learning, with higher performance on the training set than the validation and test sets. Distant supervision with the NICE CKS Health Topics and ICPC-2 Code descriptions demonstrated clear improvements. The key phrases in the ICPC-2 descriptions are a natural fit for NB: these features are individually informative, which allows linear models such as NB to perform well. The imperfect mapping between CKS topics and ICPC-2 codes may reduce the performance of NB on CKS topics. Improving the mapping would require

Table 2 Performance with standard supervised learning, 95% CIs shown in parentheses

	Validation			Train	Test		
Model	**Precision**	**Recall**	**F1**	**F1**	**Precision**	**Recall**	**F1**
Random baseline	0.499	0.104	0.161	0.161	0.501	0.102	0.158
Conventional supervision							
Naïve Bayes (multilabel)	0.284 (0.237 to 0.325)	0.140 (0.135 to 0.192)	0.175 (0.161 to 0.222)	**0.999**	0.234 (0.169 to 0.276)	0.113 (0.087 to 0.158)	0.139 (0.106 to 0.185)
Naïve Bayes (multiclass)	0.372 (0.298 to 0.399)	0.327 (0.314 to 0.398)	0.300 (0.266 to 0.342)	0.696	0.178 (0.146 to 0.232)	0.238 (0.213 to 0.294)	0.181 (0.154 to 0.226)
SVM (multilabel)	0.107 (0.112 to 0.132)	**1.000** (1.000 to 1.000)	0.184 (0.192 to 0.223)	0.181	0.102 (0.095 to 0.124)	1.000 (1.000 to 1.000)	0.177 (0.166 to 0.211)
SVM (multiclass)	0.200 (0.171 to 0.244)	0.159 (0.157 to 0.211)	0.154 (0.142 to 0.196)	0.696	0.217 (0.145 to 0.263)	0.169 (0.14 to 0.227)	0.164 (0.129 to 0.213)
Nearest centroid (multiclass)	0.349 (0.297 to 0.395)	0.270 (0.254 to 0.327)	0.278 (0.247 to 0.325)	0.694	0.307 (0.18 to 0.355)	0.205 (0.15 to 0.276)	0.219 (0.151 to 0.278)
BERT conventional (multiclass)	**0.467** (0.434 to 0.549)	0.577 (0.546 to 0.654)	**0.480** (0.447 to 0.550)	0.696	**0.484** (0.414 to 0.575)	**0.509** (0.434 to 0.610)	**0.452** (0.390 to 0.525)
Distant supervision							
Naïve Bayes (multilabel), ICPC-2	0.626 (0.515 to 0.687)	0.234 (0.196 to 0.278)	0.323 (0.268 to 0.362)	0.979	0.590 (0.427 to 0.656)	0.285 (0.206 to 0.384)	0.378 (0.274 to 0.456)
Naïve Bayes (multiclass), ICPC-2	0.516 (0.466 to 0.569)	0.590 (0.541 to 0.639)	0.512 (0.462 to 0.549)	1.00	0.511 (0.412 to 0.611)	0.524 (0.449 to 0.628)	0.481 (0.404 to 0.567)
Nearest centroid, ICPC-2	**0.718** (0.565 to 0.765)	0.416 (0.373 to 0.463)	0.444 (0.384 to 0.489)	1.00	0.520 (0.400 to 0.615)	0.362 (0.298 to 0.448)	0.386 (0.303 to 0.467)
Conventional BERT, CKS	0.603 (0.553 to 0.653)	0.584 (0.53 to 0.64)	0.550 (0.494 to 0.593)	0.927	**0.551** (0.477 to 0.649)	0.562 (0.483 to 0.691)	**0.508** (0.429 to 0.594)
BERT NSP, CKS	0.364 (0.333 to 0.394)	**0.816** (0.767 to 0.865)	0.462 (0.424 to 0.488)	0.291	0.257 (0.215 to 0.331)	**0.598** (0.525 to 0.711)	0.306 (0.257 to 0.371)
BERT MLM, CKS	0.600 (0.547 to 0.64)	0.615 (0.566 to 0.673)	**0.567** (0.512 to 0.604)	0.792	0.481 (0.409 to 0.574)	0.536 (0.469 to 0.639)	0.467 (0.397 to 0.548)

For conventional supervision, ‘train’ and ‘test’ results are for classifiers trained on the whole 80% training split, and validation was performed using 5-fold cross-validation over the training set. For distant supervision, the OIAM training set was repurposed as a validation set, as it was not used to train the models with this setup.

CKS, Clinical Knowledge Summaries; ICPC-2, International Classification of Primary Care-2; MLM, masked language modelling; NSP, next sentence prediction; OIAM, One in a Million; SVM, support vector machine.

Table 3 F1 scores for fivefold cross-validation performance on the OIAM training set with different sets of stopwords

Model	No removal	English	Medical	Custom	Medical+custom	English+custom	English+medical +custom
Naïve Bayes (multilabel)	0.157	0.159	0.154	0.143	0.166	0.170	**0.175**
Naïve Bayes (multiclass)	0.225	0.266	0.243	0.228	0.245	0.272	**0.300**
SVM (multilabel)	0.184	0.184	0.184	0.184	0.184	0.184	0.184
SVM (multiclass)	0.141	0.151	0.141	0.142	0.142	0.150	**0.154**
Nearest centroid	0.234	0.256	0.239	0.234	0.247	0.252	**0.278**

OIAM, One in a Million; SVM, support vector machine.

Table 4 F1 scores for distant supervision performance, evaluated on the OIAM training set, with different sets of stopwords, and training on either CKS topics or ICPC-2 descriptions

Model	No removal	English	Medical	Custom	Medical+custom	English+custom	English+medical+custom
NB (multilabel), ICPC-2	0.139	0.170	0.136	0.253	0.297	**0.323**	0.297
NB (multilabel), CKS	0.096	0.160	0.119	0.126	0.191	0.207	**0.234**
NB (multiclass), ICPC-2	0.324	0.354	0.307	0.461	0.471	**0.512**	0.470
NB (multiclass), CKS	0.245	0.274	0.249	0.275	0.340	0.368	**0.375**
Nearest centroid, ICPC-2	0.312	0.354	0.317	0.432	0.437	**0.445**	0.437
Nearest centroid, CKS	0.326	0.349	0.344	0.349	0.353	0.357	**0.365**

CKS, Clinical Knowledge Summaries; ICPC-2, International Classification of Primary Care-2; NB, Naïve Bayes; OIAM, One in a Million.

Table 5 F1 scores for different sources of distant supervision, and the effect of removing class A, evaluated on the OIAM training set

Model	ICPC-2	ICPC-2 without A	CKS	CKS without A	ICPC-2 and CKS combined	Combined without A
Naïve Bayes (multilabel)	0.323 (0.268, 0.362)	**0.345** (0.286, 0.389)	0.234 (0.196, 0.262)	0.249 (0.207, 0.285)	0.254 (0.21, 0.287)	0.271 (0.225, 0.308)
Naïve Bayes (multiclass)	**0.512** (0.462, 0.549)	0.508 (0.458, 0.546)	0.375 (0.325, 0.411)	0.391 (0.34, 0.428)	0.378 (0.33, 0.416)	0.385 (0.338, 0.421)
Nearest centroid	0.444 (0.384, 0.489)	0.093 (0.063, 0.12)	0.365 (0.312, 0.401)	0.086 (0.057, 0.107)	0.367 (0.315, 0.403)	0.090 (0.063, 0.113)
BERT conventional	0.057 (0.049, 0.065)	0.027 (0.02, 0.037)	0.550 (0.494, 0.593)	**0.521** (0.459, 0.565)	**0.540** (0.476, 0.576)	**0.545** (0.483, 0.590)
BERT NSP	0.285 (0.232, 0.324)	0.347 (0.309, 0.371)	0.462 (0.424, 0.488)	0.434 (0.392, 0.466)	0.445 (0.402, 0.476)	0.467 (0.425, 0.498)
BERT MLM	0.505 (0.444, 0.544)	0.486 (0.425, 0.528)	**0.567** (0.512, 0.604)	0.497 (0.441, 0.535)	0.532 (0.472, 0.571)	0.475 (0.424, 0.512)

Highest F1 scores in bold.
CKS, Clinical Knowledge Summaries; ICPC-2, International Classification of Primary Care-2; MLM, masked language modelling; NSP, next sentence prediction; OIAM, One in a Million.

costly manual editing of the scraped CKS health topics, as some CKS topics lack a one-to-one mapping to an ICPC-2 code. Still, CKS topics produce competitive performance with BERT, which was pretrained with complete sentences, suggesting that the health topics do include useful training signals. Future work could, therefore, investigate ensembles that stack[28] models trained with different sources of data.

To identify common classifier mistakes, the clinician on the research team reviewed individual consultation transcripts and their human and predicted codes and noted several distinct types of errors. First, shallow classifiers demonstrated simple linguistic errors such as misunderstanding idioms. In one consultation, the GP repeatedly mentioned 'keeping an eye on it' and the NB classifier incorrectly coded it as an ophthalmology-related consultation; BERT overcame this by avoiding reliance on isolated words as features.[29] Second, perusing specific consultations where the NLP classifier appeared to get the coding significantly wrong highlighted errors by the original human labelling team. Third, the 'A: General' category was often selected erroneously, as the class is non-specific (precision=0.154 for NB multiclass, trained on ICPC-2 descriptions), although excluding this class often hurt performance. Finally, there were examples where a lack of clinical knowledge caused errors such as the NLP classifier assuming that a consultation discussing someone's wrist was a musculoskeletal rather than a neurological issue (such as in carpal tunnel syndrome).

Many of these specific types of error relate to limitations of the dataset: its scale, labelling quality and labelling scheme; we consider its small size to be the most significant issue. When scaling up the dataset, further limitations to address include the dataset being only in English and all the consultations taking place in one part of the UK. The current areas where clinical machine learning is excelling are radiology and pathology due to their large and accessible (anonymised) datasets, and the creation of a large, anonymised, free text dataset related to primary care would be hugely valuable for research. The COVID-19 pandemic accelerated the use of online consultations producing potential sources of patient-entered free text (eg, AskMyGP[30]) and recorded audio/video consultations for examination (eg, by FourteenFish[31]). We advocate for routinely incorporating consent to use digitally recorded clinical consultations for research and providing robust anonymisation of them, so that researchers can conduct valuable and translational research in this area.

Further directions for future research include processing the consultations in 'real-time' and assigning them to the more fine-grained NICE CKS health topics rather than ICPC-2 codes, which would allow the system to link a doctor automatically to the corresponding health topic guidelines. Performance may also be improved by combining text with other data from electronic medical records.

Table 6 F1 scores when patients' transcribed speech is excluded

Model	Including GP and patient speech	Only GP speech
Naïve Bayes (multilabel) ICPC-2	0.323 (0.268, 0.362)	**0.372** (0.3, 0.417)
Naïve Bayes (multiclass) ICPC-2	**0.512** (0.462, 0.549)	0.484 (0.429, 0.521)
Nearest centroid ICPC-2	**0.444** (0.384, 0.489)	0.425 (0.361, 0.47)
BERT conventional, CKS	**0.550** (0.494, 0.593)	0.445 (0.384, 0.465)
BERT NSP, CKS	**0.462** (0.424, 0.488)	0.436 (0.398, 0.464)
BERT MLM, CKS	**0.567** (0.512, 0.604)	0.500 (0.434, 0.539)

The classifiers were trained using their most effective distant supervision source and evaluated on the OIAM training set (repurposed as a validation set). Bold indicates best performance in a comparison between including and excluding patients' speech with the same classifier. CKS, Clinical Knowledge Summaries; GP, general practitioner; ICPC-2, International Classification of Primary Care-2; MLM, masked language modelling; NSP, next sentence prediction; OIAM, One in a Million.

CONCLUSION

This paper offers a promising avenue of research using NLP to extract information from the conversation between a patient and their doctor in a primary care consultation and demonstrates a successful collaboration between clinical and computing disciplines. Previous projects using NLP in a clinical setting have focused on classifying free text created by health professionals (such as radiology reports) or generating free text from codes and defined data (such as investigation results). To our knowledge, this is the first time that the original conversation between a doctor and their patient has been analysed using NLP. Our comparison of text classifiers showed modest gains from deep learning approaches, that the models can be trained using health topics scraped from web pages, and that patients' speech contains valuable signals for assigning medical codes. We identified potential improvements, including adding colloquial vocabulary to health topic descriptions, increasing the dataset size and domain-specific pretraining of language models. Our ultimate goal would be to provide a smart digital assistant that can create effective consultation notes and suggest questions or guidelines to the clinician[32]; this is likely to require significant advances both in NLP and in our understanding of what makes good clinical notes. While this goal is still a long way off, our work represents one small step towards that reality.

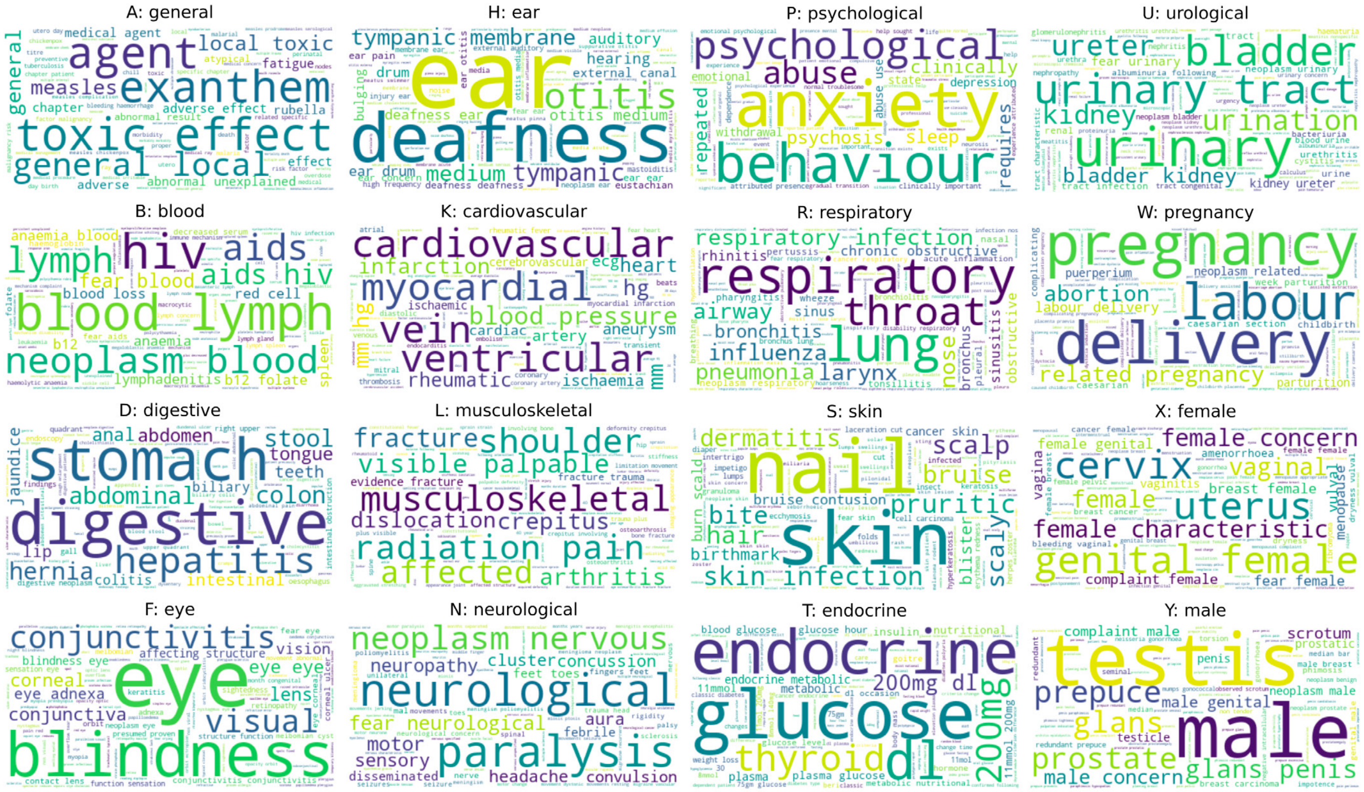


Figure 2 Wordclouds for each ICPC-2 category, with unigrams and bigrams weighted by the probability of the class label given the feature. ICPC-2, International Classification of Primary Care.

Acknowledgements This work was carried out using the computational facilities of the Advanced Computing Research Centre, University of Bristol—http://www.bristol.ac.uk/acrc/. The authors would like to thank Professor John Macleod for his support during this project.

Contributors YP and ES contributed to the conception and design of the work, the sourcing of the data and the analysis of the results. YP and ES wrote the paper with review and comments by YMW and HF. YMW, HF and ES wrote the code associated with the project. YP acts as guarantor for the work and conduct of the study, had access to the data, and controlled the decision to publish.

Funding YP is funded via an NIHR Academic Clinical Fellowship in Primary Care—Award Identifier: ACF-2018-25-007. ES is employed by the University of Bristol as a Lecturer. He has no other funding streams. YMW and HF were funded by an Elizabeth Blackwell Institute (EBI) Wellcome Grant: 204813/Z/16/Z. The 'One in a Million' study was funded by the National Institute for Health Research (NIHR) School for Primary Care Research (208) and the South West GP Trust.

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval The project operated under South West—Central Bristol Research Ethics Committee (Reference: 21/SW/0069) approval granted on 3 June 2021.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement The 'One in a Million' dataset is available for research use following valid ethics approval.ICPC-2 Codes and descriptions are freely downloadable from the web.The NICE CKS Health Topics dataset is freely downloadable from their website using freely available 'web-scraping' tools.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

ORCID iD
Yvette Pyne http://orcid.org/0000-0001-5920-484X

REFERENCES

1 Topol EJ. The topol review: preparing the healthcare workforce to deliver the digital future. 2019.
2 Young RA, Burge SK, Kumar KA, *et al*. A time-motion study of primary care physicians' work in the electronic health record era. *Fam Med* 2018;50:91–9.
3 Sinsky C, Colligan L, Li L, *et al*. Allocation of physician time in ambulatory practice: a time and motion study in 4 specialties. *Ann Intern Med* 2016;165:753–60.
4 Yim W-W, Yetisgen M, Harris WP, *et al*. Natural language processing in oncology: A review. *JAMA Oncol* 2016;2:797–804.
5 Khanbhai M, Anyadi P, Symons J, *et al*. Applying natural language processing and machine learning techniques to patient experience feedback: a systematic review. *BMJ Health Care Inform* 2021;28:e100262.
6 Khalifa A, Meystre S. Adapting existing natural language processing resources for cardiovascular risk factors identification in clinical notes. *J Biomed Inform* 2015;58 Suppl(Suppl):S128–32.
7 Koleck TA, Dreisbach C, Bourne PE, *et al*. Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review. *J Am Med Inform Assoc* 2019;26:364–79.
8 Doan S, Maehara CK, Chaparro JD, *et al*. Building a natural language processing tool to identify patients with high clinical suspicion for kawasaki disease from emergency department notes. *Acad Emerg Med* 2016;23:628–36.
9 Moen H, Peltonen L-M, Heimonen J, *et al*. Comparison of automatic summarisation methods for clinical free text notes. *Artif Intell Med* 2016;67:25–37.
10 Corrente M, Bourgeault I. *Innovation in transcribing data: meet otter.ai*. 1 Oliver's Yard, 55 City Road, London EC1Y 1SP United Kingdom, 2022.
11 How robin works. Available: https://www.robinhealthcare.com/how-robin-works [Accessed 14 Mar 2023].
12 Quiroz JC, Laranjo L, Kocaballi AB, *et al*. Challenges of developing a digital scribe to reduce clinical documentation burden. *NPJ Digit Med* 2019;2:114.
13 van Buchem MM, Boosman H, Bauer MP, *et al*. The digital scribe in clinical practice: a scoping review and research agenda. *NPJ Digit Med* 2021;4:57.
14 Krishna K, Pavel A, Schloss B, *et al*. Extracting structured data from physician-patient conversations by predicting noteworthy utterances. 2020.
15 Barnes R. One in A million: A study of primary care consultations. 2017.
16 Jepson M, Salisbury C, Ridd MJ, *et al*. The "one in a million" study: creating a database of uk primary care consultations. *Br J Gen Pract* 2017;67:e345–51.
17 World Organization of National Colleges A and Academic Associations of General Practitioners, Family Physicians, Classification Committee. *International classification of primary care: ICPC-2*. Oxford: Oxford Univ. Press, 1998.
18 ICPC-2e – english version. ehelse. Available: https://www.ehelse.no/kodeverk/icpc-2e--english-version [Accessed 5 May 2022].
19 NICE. Health topics A to Z | CKS | NICE. Available: https://cks.nice.org.uk/topics/ [Accessed 11 Feb 2022].
20 Bobo WV, Pathak J, Kremers HM, *et al*. An electronic health record driven algorithm to identify incident antidepressant medication users. *J Am Med Inform Assoc* 2014;21:785–91.
21 Gu Y, Tinn R, Cheng H, *et al*. Domain-Specific language model pretraining for biomedical natural language processing. *ACM Trans Comput Healthcare* 2022;3:1–23.
22 Devlin J, Chang M-W, Lee K, *et al*. BERT: pre-training of deep bidirectional transformers for language understanding. 2021. Available: http://arxiv.org/abs/1810.04805
23 Radford A, Wu J, Child R, *et al*. n.d. Language models are unsupervised multitask learners. ;24.
24 Schick T, Schütze H. Exploiting cloze-questions for few-shot text classification and natural language inference. In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2021: 255–69.
25 Qu J, Hashimoto K, Liu W, *et al*. Few-shot intent classification by gauging entailment relationship between utterance and semantic label. Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI; Stroudsburg, PA, USA: Association for Computational Linguistics, 2021:8–15
26 Lee J, Yoon W, Kim S, *et al*. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 2020;36:1234–40.
27 Huang K, Altosaar J, Ranganath R. ClinicalBERT: modeling clinical notes and predicting hospital readmission. 2020. Available: http://arxiv.org/abs/1904.05342
28 Wolpert DH. Stacked generalization. *Neural Networks* 1992;5:241–59.
29 Arora S, May A, Zhang J, *et al*. Contextual embeddings: when are they worth it? Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics; Online. Stroudsburg, PA, USA: Association for Computational Linguistics, 2020:2650–63
30 AskmyGP | the most effective online triage and consultation tool for gps. askmyGP. Available: https://askmygp.uk/ [Accessed 8 Jul 2022].
31 Appraisal Toolkit, AKT, RCA, Education, Trainee Portfolio - FourteenFish. Trainee portfolio - fourteenfish. Available: https://www.fourteenfish.com/ [Accessed 8 Jul 2022].
32 Stewart S, Pyne Y, McMillan B. Augmented consulting: the future of primary care? *BJGP Open* 2021;5:BJGPO.2020.0177.