

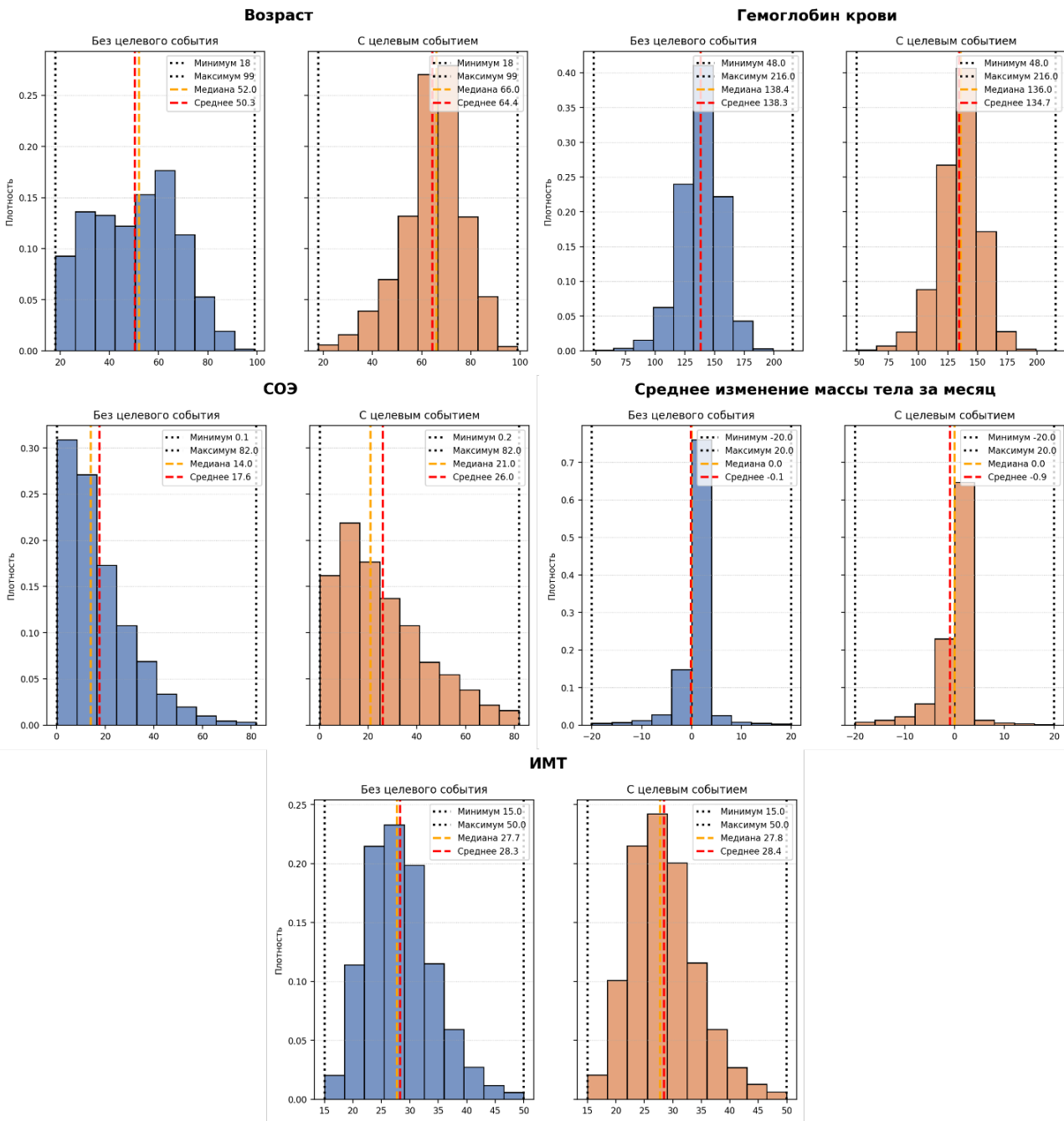
Приложение 1. Описательная статистика. Результаты анализа количественных данных представлены в виде медианы (Me) с указанием межквартильного размаха (IQR), а категориальных — в виде долей

Признак	Без целевого события в данных для разработки (N = 1 070 762)	С целевым событием в данных для разработки (N = 167 995)	p	Набор для разработки (N = 1 238 757)	Внешний набор данных № 1 (N = 29 681)	p	Внешний набор данных № 2 (N = 25 145)	p
Среднее изменение массы тела за месяц	0,0 (0,0, 0,0)	0,0 (-1,0, 0,0)	< 0,05	0,0 (0,0, 0,0)	0,0 (-0,5, 0,3)	< 0,05	0,0 (-0,7, 0,0)	< 0,05
Возраст	52,0 (36,0, 63,0)	66,0 (58,0, 73,0)	< 0,05	55,0 (38,0, 65,0)	53,0 (36,0, 65,0)	< 0,05	55,0 (37,0, 65,0)	0,595
Гемоглобин крови	139,0 (128,0, 149,0)	136,0 (125,0, 146,0)	< 0,05	138,0 (128,0, 149,0)	127,0 (115,0, 139,0)	< 0,05	132,0 (122,0, 142,7)	< 0,05
Индекс массы тела	27,8 (24,1, 32,0)	27,8 (24,2, 32,0)	0,073	27,8 (24,1, 32,0)	26,7 (23,0, 30,5)	< 0,05	27,7 (24,0, 31,6)	< 0,05
Мужской пол	361 562 (33,8 %)	65 656 (39,1 %)	< 0,05	427 218 (34,5 %)	8606 (29,0 %)	< 0,05	7543 (30,0 %)	< 0,05
Клинически значимое сопутствующее заболевание	172 752 (16,1 %)	58 122 (34,6 %)	< 0,05	230874 (18,6 %)	2843 (9,6 %)	< 0,05	5964 (23,7 %)	< 0,05
COЭ	14,0 (7,0, 24,0)	22,0 (12,0, 37,0)	< 0,05	15,0 (7,0, 26,0)	18,0 (10,0, 30,0)	< 0,05	16,0 (9,0, 28,0)	< 0,05

Appendix 1. Descriptive statistics of the collected data. The results of the quantitative data analysis are presented as the median (Me) with the interquartile range (IQR), while categorical data are presented as proportions

Feature	Without the target event in the development data (N = 1,070,762)	With the target event in the development data (N = 167,995)	p	Development set (N = 1,238,757)	External data set No 1 (N = 29,681)	p	External data set No 2 (N = 25,145)	p
Mean monthly weight change rate	0.0 (0.0, 0.0)	0.0 (-1.0, 0.0)	< 0.05	0.0 (0.0, 0.0)	0.0 (-0.5, 0.3)	< 0.05	0.0 (-0.7, 0.0)	< 0.05
Age	52.0 (36.0, 63.0)	66.0 (58.0, 73.0)	< 0.05	55.0 (38.0, 65.0)	53.0 (36.0, 65.0)	< 0.05	55.0 (37.0, 65.0)	0.595
Hemoglobin	139.0 (128.0, 149.0)	136.0 (125.0, 146.0)	< 0.05	138.0 (128.0, 149.0)	127.0 (115.0, 139.0)	< 0.05	132.0 (122.0, 142.7)	< 0.05
BMI	27.8 (24.1, 32.0)	27.8 (24.2, 32.0)	0.073	27.8 (24.1, 32.0)	26.7 (23.0, 30.5)	< 0.05	27.7 (24.0, 31.6)	< 0.05
Gender (male)	361,562 (33.8 %)	65,656 (39.1 %)	< 0.05	427,218 (34.5 %)	8606 (29.0 %)	< 0.05	7543 (30.0 %)	< 0.05
Clinically significant comorbidity	172752 (16.1 %)	58122 (34.6 %)	< 0.05	230,874 (18.6 %)	2843 (9.6 %)	< 0.05	5964 (23.7 %)	< 0.05
ESR	14.0 (7.0, 24.0)	22.0 (12.0, 37.0)	< 0.05	15.0 (7.0, 26.0)	18.0 (10.0, 30.0)	< 0.05	16.0 (9.0, 28.0)	< 0.05

Приложение 2. Распределение значений количественных признаков в собранном наборе данных
Appendix 2. Distribution of quantitative feature values in the collected data



Приложение 3. Коды по МКБ-10, при наличии которых в листе окончательных диагнозов пациента значение признака «Клинически значимое сопутствующее заболевание» становилось True

№	Клинически значимое сопутствующее заболевание	Код МКБ-10
1	Кератодермия	L85.1
2	Грыжи пищевода	K44
3	Синдром Пейтца — Егерса	Q85.8
4	Аденоматозные полипы кишечника	K63.5
5	Аденоматозные полипы тела матки	D28
6	Болезнь Боуэна	D04. 9
7	Эритроплазия Кейра	D07.4
8	Язва желудка	K25
9	Семейный полипоз толстой кишки	D12.6
10	Полипы желудка и двенадцатиперстной кишки	K31.7
11	Полип прямой кишки	D12.8, K62.1
12	Ахалазия кардиальной части пищевода	K44
13	Пищевод Барретта	K22.7
14	Множественный эндокринный аденоматоз	E34.8, D13.7, D35.0–D35.2, D35.8
15	Множественная эндокринная неоплазия: тип 2А (синдром Сиппла); тип 2В (синдром Горлина)	D44.8, D35.0, D35.1 D35.8
16	Заболевания с нарушением формирования пола	E34.5
17	Акромегалия	E22.0
18	Нетоксический зоб	E04.1, E04.2
19	Тиреотоксикоз с токсическим зобом	E05.1, E05.2
20	Аденома паращитовидной железы	D35.1
21	Первичный гиперпаратиреоз	E21.0
22	Аденома надпочечника	D35.0
23	Туберозный склероз	Q85.1
24	Доброкачественное новообразование больших слюнных желез	D11
25	Полиостозная фиброзная дисплазия	Q78.1
26	Папилломы, полипы мочевого пузыря	D30.3
27	Папилломы, полипы мочеиспускательного канала	D30.4
28	Лейкоплакия полового члена	N 48.0
29	Сложные кисты почки	D41.0
30	Ангиомиолипома почки	D30.0
31	Простатическая интраэпителиальная неоплазия простаты	D29.1
32	Поражения костно-мышечной системы после медицинских процедур на опорно-двигательном аппарате в связи с опухолевым и системным поражением	M96
33	Болезнь Педжета	M88
34	Солитарные и множественные остеохондромы	D16
35	Фиброзная дисплазия	M85
36	Энхондроматоз	Q78.4
37	Доброкачественное новообразование глаза и его придаточного аппарата	D31
38	Доброкачественные новообразования кожи века, включая спайку век	D23.1
39	Полип голосовой складки и гортани	J38.1
40	Папилломатоз, фиброматоз гортани	D14.1
41	Доброкачественное новообразование трахеи	D14.2
42	Инвертированная папиллома полости носа	D14.0
43	Полип носа	J33
44	Новообразование среднего уха	D14
45	Доброкачественное новообразование миндалин	D10.4
46	Доброкачественное новообразование других частей ротоглотки	D10.5
47	Доброкачественное новообразование носоглотки	D10.6
48	Доброкачественное новообразование гортаноглотки	D10.7

№	Клинически значимое сопутствующее заболевание	Код МКБ-10
49	Доброкачественное новообразование глотки неуточненной локализации	D10.9
50	Лейкоплакия и другие изменения эпителия полости рта, включая язык	K13.2
51	Абразивный хейлит Манганотти, ограниченный гиперкератоз, бородавчатый предрак	K13.0
52	Доброкачественное новообразование губы	D10.0
53	Доброкачественное новообразование языка	D10.1
54	Доброкачественное новообразование дна полости рта	D10.2
55	Доброкачественное новообразование других неуточненных частей рта	D10.3
56	Меланоз полости рта	K13.7
57	Полиостозная фиброзная дисплазия	Q78.1
58	Красный плоский лишай	L43
59	Синдром диспластических невусов, синдром FAMM, FAMMM	D22
60	Врожденные гигантские и крупные невусы	Q82.5
61	Невус Ядассона, синдром Горлина — Гольца, синдром Базекса, синдром Рембо	D23
62	Актинический кератоз	L57.1
63	Эруптивный себорейный кератоз	L82
64	Ксеродерма пигментная	Q82.1
65	Полипы шейки матки и эндометрия	N84
66	Синдром поликистоза яичников	E28.2
67	Лейкоплакия шейки матки	N88.0
68	Железистая гиперплазия эндометрия	N85.0
69	Аденоматозная гиперплазия эндометрия	N85.1
70	Умеренная дисплазия шейки матки	N87.1
71	Резко выраженная дисплазия шейки матки, не классифицированная в других рубриках	N87.2
72	Новообразования неопределенного или неизвестного характера яичника	D39.1
73	Доброкачественное новообразование молочной железы	D24
74	Доброкачественная дисплазия молочной железы	N60

Appendix 3. ICD-10 codes defining clinically significant comorbidities. Patients with any of these codes in their final diagnosis records were classified as having clinically significant comorbidities (value = True)

No	Clinically Significant Comorbid Condition	ICD-10 code
1	Keratoderma	L85.1
2	Esophageal Hernias	K44
3	Peutz-Jeghers Syndrome	Q85.8
4	Adenomatous Polyps of the Intestine	K63.5
5	Adenomatous Polyps of the Uterine Body	D28
6	Bowen's Disease	D04.9
7	Keir's Erythroplasia	D07.4
8	Stomach Ulcer	K25
9	Familial Adenomatous Polyposis of the Colon	D12.6
10	Polyps of the Stomach and Duodenum	K31.7
11	Rectal Polyp	D12.8, K62.1
12	Achalasia of the Cardiac Part of the Esophagus	K44
13	Barrett's Esophagus	K22.7
14	Multiple Endocrine Adenomatosis	E34.8, D13.7, D35.0–D35.2, D35.8
15	Multiple Endocrine Neoplasia: Type 2A (Sipple Syndrome); Type 2B (Gorlin Syndrome)	D44.8, D35.0, D35.1 D35.8
16	Disorders with Sex Development Abnormalities	E34.5
17	Acromegaly	E22.0
18	Nontoxic Goiter	E04.1, E04.2
19	Thyrotoxicosis with Toxic Goiter	E05.1, E05.2
20	Parathyroid Adenoma	D35.1
21	Primary Hyperparathyroidism	E21.0

No	Clinically Significant Comorbid Condition	ICD-10 code
22	Adrenal Adenoma	D35.0
23	Tuberous Sclerosis	Q85.1
24	Benign Tumors of the Large Salivary Glands	D11
25	Polyostotic Fibrous Dysplasia	Q78.1
26	Papillomas, Polyps of the Urinary Bladder	D30.3
27	Papillomas, Polyps of the Urethra	D30.4
28	Leukoplakia of the Penis	N 48.0
29	Complex Kidney Cysts	D41.0
30	Angiomyolipoma of the Kidney	D30.0
31	Prostatic Intraepithelial Neoplasia of the Prostate	D29.1
32	Bone-Muscle System Lesions Following Medical Procedures on the Musculoskeletal System Due to Tumor and Systemic Involvement	M96
33	Paget's Disease	M88
34	Solitary and Multiple Osteochondromas	D16
35	Fibrous Dysplasia	M85
36	Enchondromatosis	Q78.4
37	Benign Tumors of the Eye and its Adnexal Apparatus	D31
38	Benign Tumors of the Eyelids, Including Eyelid Adhesions	D23.1
39	Polyp of the Vocal Cord and Larynx	J38.1
40	Papillomatosis, Fibromatosis of the Larynx	D14.1
41	Benign Tumors of the Trachea	D14.2
42	Inverted Papilloma of the Nasal Cavity	D14.0
43	Nasal Polyp	J33
44	Middle Ear Neoplasm	D14
45	Benign Tumor of the Tonsil	D10.4
46	Benign Tumor of Other Parts of the Oropharynx	D10.5
47	Benign Tumor of the Nasopharynx	D10.6
48	Benign Tumor of the Laryngopharynx	D10.7
49	Benign Tumor of the Pharynx of Unspecified Localization	D10.9
50	Leukoplakia and Other Epithelial Changes of the Oral Cavity, Including the Tongue	K13.2
51	Abrasive Cheilitis of Manganotti, Limited Hyperkeratosis, Verrucous Precancer	K13.0
52	Benign Tumor of the Lip	D10.0
53	Benign Tumor of the Tongue	D10.1
54	Benign Tumor of the Floor of the Mouth	D10.2
55	Benign Tumor of Other Unspecified Parts of the Mouth	D10.3
56	Melanosis of the Oral Cavity	K13.7
57	Polyostotic Fibrous Dysplasia (Repeated)	Q78.1
58	Lichen Planus	L43
59	Dysplastic Nevus Syndrome, FAMM Syndrome, FAMMM Syndrome	D22
60	Congenital Giant and Large Nevi	Q82.5
61	Jadassohn Nevus, Gorlin-Goltz Syndrome, Bazex Syndrome, Rembo Syndrome	D23
62	Actinic Keratosis	L57.1
63	Eruptive Seborrheic Keratosis	L82
64	Xeroderma Pigmentosum	Q82.1
65	Polyps of the Cervix and Endometrium	N 84
66	Polycystic Ovary Syndrome	E28.2
67	Leukoplakia of the Cervix	N88.0
68	Glandular Hyperplasia of the Endometrium	N85.0
69	Adenomatous Hyperplasia of the Endometrium	N85.1
70	Moderate Cervical Dysplasia	N87.1
71	Severe Cervical Dysplasia Not Classified in Other Categories	N87.2
72	Neoplasms of Unspecified or Unknown Character of the Ovary	D39.1
73	Benign Tumors of the Breast	D24
74	Benign Breast Dysplasia	N60

Приложение 4. Метрики качества итоговой модели при проведении внутреннего тестирования и внешней валидации с использованием различных порогов бинаризации

Набор дан-ных	Тип	Значе-ние	ИО	ЛП	ЛО	ИП	AUROC	AUPRC	Точ-ность	Чув-ствитель-ность	Спец-ифич-ность	Сбалан-сированная точность	ПЦПР	ПЦОР	LR+	LR-	Brier
Данные для внутреннего тестирования (247 752)	Индекс Юдена	0,131	146 636	67 555	7751	25 810	0,807 [0,798–0,815]	0,417 [0,400–0,434]	0,696 [0,690–0,703]	0,769 [0,753–0,785]	0,685 [0,677–0,692]	0,727 [0,718–0,735]	0,276 [0,271–0,282]	0,950 [0,946–0,953]	2,438 [2,369–2,513]	0,337 [0,314–0,361]	0,097 [0,096–0,099]
	Порог для це-левой ПЦПР	0,391	204 360	9831	23 731	9830			0,865 [0,861–0,868]	0,293 [0,276–0,310]	0,954 [0,951–0,957]	0,624 [0,615–0,632]	0,500 [0,478–0,522]	0,896 [0,894–0,898]	6,381 [5,849–6,963]	0,741 [0,723–0,759]	
	Порог для це-левой ПЦОР	0,021	49 726	164 465	503	33 058			0,334 [0,329–0,339]	0,985 [0,980–0,990]	0,232 [0,226–0,238]	0,609 [0,605–0,612]	0,167 [0,166–0,169]	0,990 [0,987–0,993]	1,283 [1,271–1,295]	0,065 [0,045–0,085]	
	Индекс Юдена	0,131	15 139	9155	1060	4327			0,656 [0,649–0,662]	0,803 [0,790–0,816]	0,623 [0,616–0,631]	0,713 [0,706–0,720]	0,321 [0,316–0,327]	0,935 [0,930–0,938]	2,131 [2,079–2,187]	0,316 [0,296–0,337]	
Внешн, набор дан-ных № 1 (29 681)	Порог для це-левой ПЦПР	0,391	22 741	1553	3666	1721	0,794 [0,786–0,801]	0,466 [0,451–0,482]	0,824 [0,820–0,828]	0,319 [0,303–0,334]	0,936 [0,932–0,940]	0,628 [0,620–0,635]	0,526 [0,507–0,543]	0,861 [0,858–0,864]	4,998 [4,633–5,369]	0,727 [0,711–0,745]	0,123 [0,122–0,125]
	Порог для це-левой ПЦОР	0,021	5385	18 909	69	5318			0,361 [0,355–0,366]	0,987 [0,983–0,991]	0,222 [0,215–0,228]	0,604 [0,601–0,608]	0,220 [0,218–0,221]	0,987 [0,984–0,991]	1,268 [1,257–1,280]	0,058 [0,042–0,075]	
	Индекс Юдена	0,131	11 761	9382	617	3385			0,602 [0,596–0,609]	0,846 [0,833–0,859]	0,556 [0,549–0,564]	0,701 [0,693–0,708]	0,265 [0,261–0,269]	0,950 [0,946–0,954]	1,906 [1,862–1,949]	0,277 [0,254–0,301]	
Внешний на-бор данных № 2 (25 145)	Порог для це-левой ПЦПР	0,391	19 257	1886	2442	1560	0,790 [0,782–0,798]	0,435 [0,419–0,451]	0,828 [0,824–0,832]	0,390 [0,374–0,406]	0,911 [0,907–0,915]	0,650 [0,642–0,658]	0,453 [0,437–0,468]	0,887 [0,885–0,890]	4,370 [4,103–4,650]	0,670 [0,652–0,688]	0,114 [0,112–0,115]
	Порог для це-левой ПЦОР	0,021	4017	17126	41	3961			0,317 [0,312–0,322]	0,990 [0,986–0,993]	0,190 [0,184–0,196]	0,590 [0,587–0,593]	0,188 [0,187–0,189]	0,99 [0,986–0,993]	1,222 [1,212–1,231]	0,054 [0,036–0,073]	

Примечание: ИО — истинно отрицательные, ЛП — ложноположительные, ЛО — ложноотрицательные, ИП — истинно положительные, ПЦПР — прогностическая ценность положительного результата, ПЦОР — прогностическая ценность отрицательного результата.

Appendix 4. Quality metrics of the final model during testing and external validation using different activation thresholds

Dataset	Thresholds' type	Value	TN	FP	FN	TP	AUROC	AUPRC	Acc.	Sens.	Spec.	Bal-anced Acc.	PPV	NPV	LR+	LR-	Brier
Internal test set (247, 752)	Youden's Index	0.131	146,636	67,555	7751	25,810			0.696 [0.690–0.703]	0.769 [0.753–0.785]	0.685 [0.677–0.692]	0.727 [0.718–0.735]	0.276 [0.271–0.282]	0.950 [0.946–0.953]	2.438 [2.369–2.513]	0.337 [0.314–0.361]	
	Threshold for Target PPV	0.391	204,360	9831	23,731	9830	0.807 [0.798–0.815]	0.417 [0.400–0.434]	0.865 [0.861–0.868]	0.293 [0.276–0.310]	0.954 [0.951–0.957]	0.624 [0.615–0.632]	0.500 [0.478–0.522]	0.896 [0.894–0.898]	6.381 [5.849–6.963]	0.741 [0.723–0.759]	0.097 [0.096–0.099]
	Threshold for Target NPV	0.021	49,726	164,465	503	33,058			0.334 [0.329–0.339]	0.985 [0.980–0.990]	0.232 [0.226–0.238]	0.609 [0.605–0.612]	0.167 [0.166–0.169]	0.990 [0.987–0.993]	1.283 [1.271–1.295]	0.065 [0.045–0.085]	
	Youden's Index	0.131	15,139	9155	1060	4327			0.656 [0.649–0.662]	0.803 [0.790–0.816]	0.623 [0.616–0.631]	0.713 [0.706–0.720]	0.321 [0.316–0.327]	0.935 [0.930–0.938]	2.131 [2.079–2.187]	0.316 [0.296–0.337]	
External data set No 1 (N = 29,681)	Threshold for Target PPV	0.391	22,741	1553	3666	1721	0.794 [0.786–0.801]	0.466 [0.451–0.482]	0.824 [0.820–0.828]	0.319 [0.303–0.334]	0.936 [0.932–0.940]	0.628 [0.620–0.635]	0.526 [0.507–0.543]	0.861 [0.858–0.864]	4.998 [4.633–5.369]	0.727 [0.711–0.745]	0.123 [0.122–0.125]
	Threshold for Target NPV	0.021	5385	18,909	69	5318			0.361 [0.355–0.366]	0.987 [0.983–0.991]	0.222 [0.215–0.228]	0.604 [0.601–0.608]	0.220 [0.218–0.221]	0.987 [0.984–0.991]	1.268 [1.257–1.280]	0.058 [0.042–0.075]	
	Youden's Index	0.131	11,761	9382	617	3385			0.602 [0.596–0.609]	0.846 [0.833–0.859]	0.556 [0.549–0.564]	0.701 [0.693–0.708]	0.265 [0.261–0.269]	0.950 [0.946–0.954]	1.906 [1.862–1.949]	0.277 [0.254–0.301]	
External data set No 2 (N = 25,145)	Threshold for Target PPV	0.391	19,257	1886	2442	1560	0.790 [0.782–0.798]	0.435 [0.419–0.451]	0.828 [0.824–0.832]	0.390 [0.374–0.406]	0.911 [0.907–0.915]	0.650 [0.642–0.658]	0.453 [0.437–0.468]	0.887 [0.885–0.890]	4.370 [4.103–4.650]	0.670 [0.652–0.688]	0.114 [0.112–0.115]
	Threshold for Target NPV	0.021	4017	17,126	41	3961			0.317 [0.312–0.322]	0.990 [0.986–0.993]	0.190 [0.184–0.196]	0.590 [0.587–0.593]	0.188 [0.187–0.189]	0.99 [0.986–0.993]	1.222 [1.212–1.231]	0.054 [0.036–0.073]	

Legend: TN — True Negatives, FP — False Positives, FN — False Negatives, TP — True Positives, Acc. — accuracy, Sens. — sensitivity, Spec. — specificity, PPV — positive predictive value, NPV — negative predictive value, LR — likelihood ratio.

Приложение 5. Примеры интерпретации работы модели. А — запись с самым высоким риском, Б — запись с самым низким риском в наборе данных для внутреннего тестирования.

Красным цветом указаны значения Шепли-векторов, увеличивающих вероятность целевого события, синим — уменьшающих ее, $f(x)$ — сумма всех Шепли-векторов для отдельной записи, $E[f(x)]$ — среднее значение суммы векторов для всех записей в наборе данных для внутреннего тестирования

Appendix 5. Model interpretation examples: (A) highest-risk case and (Б) lowest-risk case from the internal validation dataset. Shapley vector values increasing prediction probability are shown in red; decreasing probability in blue. Notation: $f(x)$ = sum of all Shapley values for an individual record; $E[f(x)]$ = expected value across all validation records

