Математический метод предварительной дифференциальной диагностики респираторных вирусных инфекций

A Preprint

Полина Юрьевна Кривуля
Факультет ВМК
МГУ имени М.В.Ломоносова
Москва, Россия
polina_krivulya@mail.ru

Олег Валентинович Сенько
Факультет ВМК
МГУ имени М.В.Ломоносова
Москва, Россия
senkoov@mail.ru

Abstract

В настоящей исследовательской работе представлены анализ и оценка характеристик нескольких респираторных вирусных инфекций: грипп, аденовирус, респираторно-синцитиальный вирус и бокавирус. В рамках данного исследования проведен анализ набора симптомов, которые свойственны каждому из них.

Оообенность задачи заключается в отсутствии табличной обучающей выборки, все признаки заданы с помощью статей с описаниями частоты встречаемости. Для распознавания конкретного типа вируса использованы математические методы, основанные на статистических данных по симптомам, полученных из различных статей.

Преимуществом данного подхода является его актуальность в случаях, когда наличие полноценной обучающей выборки невозможно.

Keywords Машинное обучение · Статистический анализ · Эпидемиология

1 Introduction

Острые респираторные вирусные инфекции (ОРВИ) — наиболее распространенная патология, на долю которой приходится около 90% всех инфекционных болезней. По данным Всемирной организации здравоохранения (ВОЗ), ежегодно заболевает каждый третий житель планеты (1) (2) (3) (4).

В Российской Федерации ежегодно болеют ОРВИ и гриппом более 30 млн человек, а суммарный экономический ущерб от респираторных вирусных инфекций оценивается в 40 млрд руб. в год. Большие экономические потери обусловлены вовлечением в эпидемический процесс трудоспособного населения, развитием осложнений, непродолжительным нестойким иммунитетом, определяющим повторные случаи заболевания (4) (5) (6) (7) (8).

Вирусы, вызывающие респираторные инфекции, не являются эндемичными для какого-либо региона или страны и распространены повсеместно. Чаще эпидемии возникают в зимнее время, однако вспышки наблюдаются и в осенне-весенний период, а спорадические случаи ОРВИ — круглый год. Известно около 300 возбудителей респираторных инфекций, более 200 из них представители 4 семейств РНК-содержащих вирусов (ортомиксовирусы, парамиксовирусы, коронавирусы и пикорнавирусы) и 2 семейств ДНК-содержащих вирусов (аденовирусы и герпесвирусы) (4) (9).

Цель данного исследования заключается в анализе и оценке характеристик нескольких типов вирусов, а именно гриппа, аденовируса, респираторно-синцитиального вируса (РС) и бокавируса. Для достижения данной цели проведен анализ набора симптомов, характерных для каждого из этих вирусов, а также учтены сезонные особенности. Ключевыми этапами данного исследования являются оценка статистической значимости симптомов и вычисление априорных и апостериорных вероятностей наличия каждого типа вируса. Для дальнейшего распознавания типа вируса по набору признаков использован

алгоритм наивного байесовского классификатора. Результаты данного исследования могут быть полезны для лучшего понимания характеристик и распространения данных вирусов, а также для разработки эффективных методов профилактики и лечения.

2 Постановка задачи

В данной работе использованы четыре таблицы с информацией о характеристиках гриппа, аденовируса, респираторно-синцитиального вируса и бокавируса. Симптомы, рассматриваемые в данном исследовании, включают температуру, кашель, одышку, боли в горле, боль в груди, ринит, головную боль, диарею и боли в животе. Информация о суммарном числе заболевших и количестве людей, имеющих каждый из симптомов, была собрана из статей. В некоторых статьях отсутствует информация о наличии части симптомов у пациентов. Отсутствует информация о симптоме "Боль в груди" у аденовируса во всех статьях, а также "Боль в груди" и "Головная боль" у бокавируса.

Для получения априорной вероятности наличия каждого типа вируса по сезонам использована отдельная таблица, отражающая заболеваемость каждым из вирусов по датам.

Для оценки качества используется метрика ROC-AUC, как рекомендуемая Министерством Здравоохранения для использования в доказательной медицине.

3 Анализ вирусов

3.1 Априорные вероятности

Суммарное количество заболеваний каждым из вирусов представлено на гистограмме 1. На рисунках 2, 3 приведены графики заболеваний вирусами по месяцам.

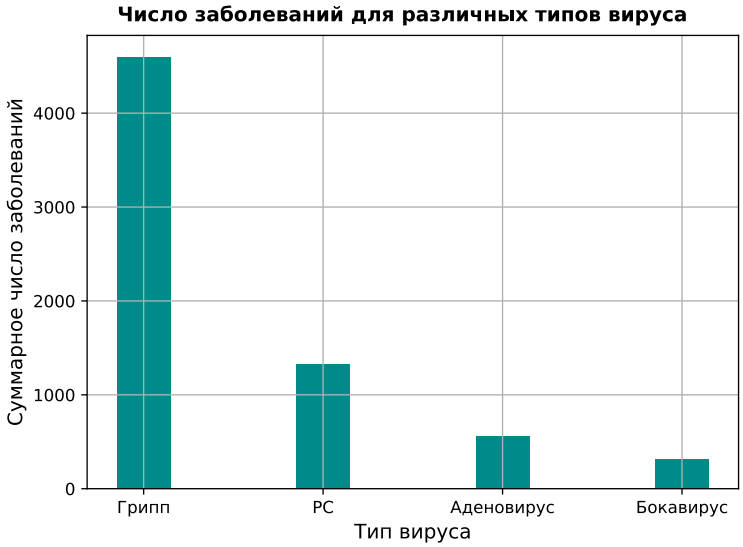

Рис. 1: Информация о суммарном числе заболевших каждым типом вируса

Выполним подсчет априорных вероятностей с использованием формулы

$$p_i^a = \frac{n_i}{\sum_{j=1}^{4} n_j}, i = \overline{1,4},$$

где i – номер типа вируса, n_i – число заболеваний i-м типом вируса в рассматриваемый сезон.

Подсчет выполняется как с учетом сезона, так и без. Полученные результаты представлены в таблице 1 и позволяют определить зависимость априорной вероятности каждого типа вируса от сезона.

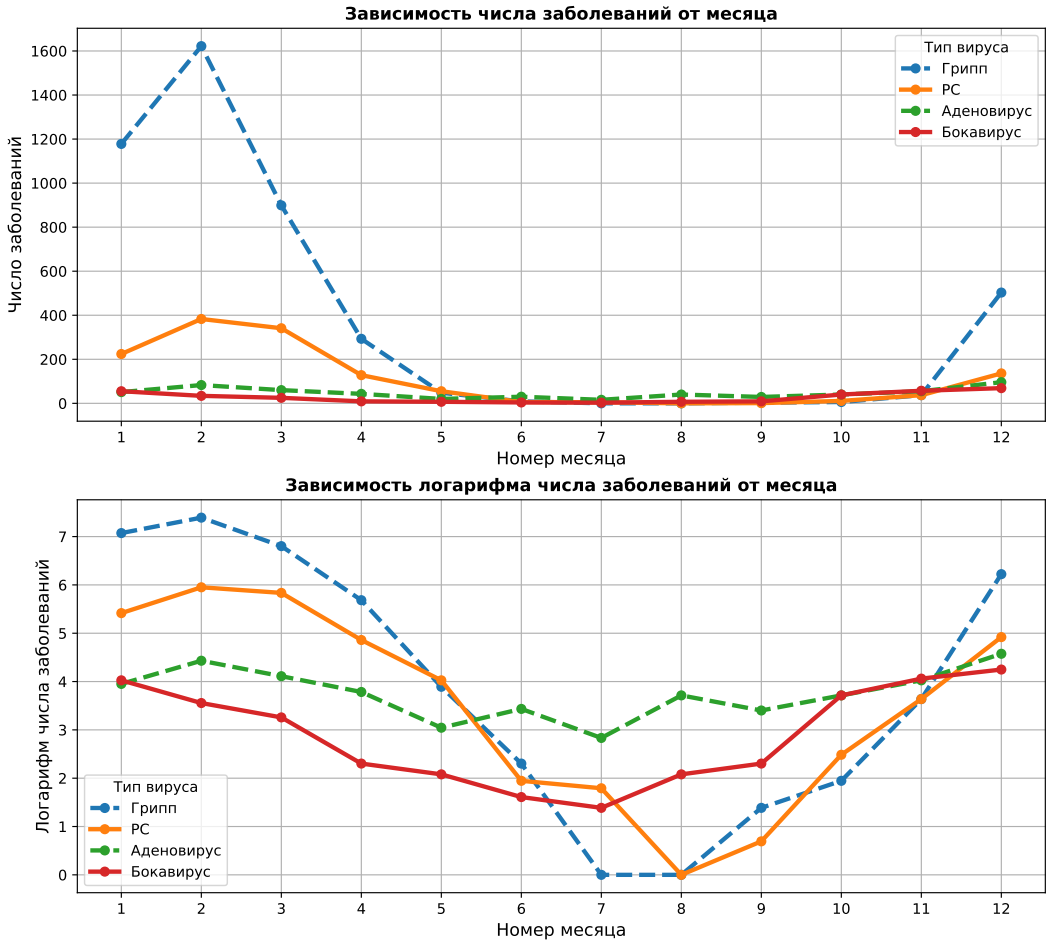

Рис. 2: Информация о числе заболевших данными типами вирусов по месяцам в обычном и логарифмическом масштабе

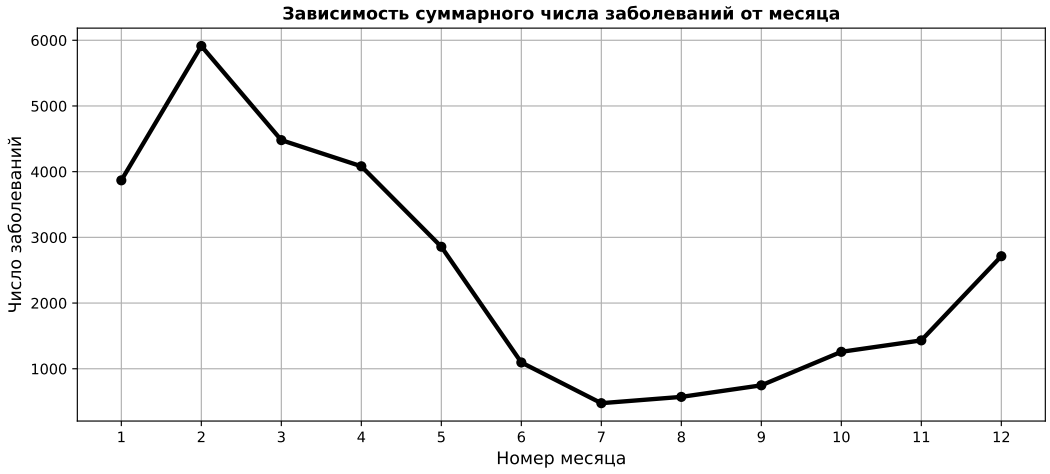

Рис. 3: Информация о суммарном числе заболевших по месяцам

Сезон \ Тип вируса	Грипп	РС	Адено	Бока
Зима	0.745	0.168	0.052	0.037
Весна	0.643	0.272	0.064	0.021
Лето	0.075	0.092	0.717	0.117
Осень	0.142	0.151	0.382	0.326
Без учета сезона	0.675	0.195	0.083	0.047

Таблица 1: Зависимость априорной вероятности каждого типа вируса от сезона

3.2 Апостериорные вероятности

Симптом \ Тип вируса	Грипп	РС	Адено	Бока
Температура	4200	698	1791	308
Кашель	3952	964	1518	373
Одышка	1382	656	172	228
Боли в горле	858	37	1204	1
Боль в груди	353	46	–	–
Ринит	1092	535	915	99
Головная боль	283	2	115	–
Диарея	221	65	362	81
Боли в животе	111	14	188	59
Число заболеваний	5957	1186	1999	525

Таблица 2: Информация о количестве зафиксированных симптомов для каждого из заболеваний (как суммарное по статьям)

Для каждого заболевания определено количество зафиксированных симптомов и суммарное число заболеваний (по всем статьям для рассматриваемого типа вируса). Эти данные представлены в таблице 2.

Для каждого типа вируса найдем вероятность иметь один из девяти симптомов как среднее по статьям в соответствии с формулой:

$$p_j^i = \frac{1}{N^i} \sum_{k=1}^{N^i} \nu_{jk}^i,$$

где ν_{jk}^i – доля симптома j в статье k для вируса i; N^i – число статей с информаией для i-го вируса.

При этом в подсчете такого среднего не учитываются статьи, где данный симптом не упомянут.

Полученные результаты представлены в таблице 3.

Симптом \ Тип вируса	Грипп	РС	Адено	Бока
Температура	0.789	0.716	0.892	0.615
Кашель	0.744	0.874	0.844	0.918
Одышка	0.400	0.601	0.228	0.557
Боли в горле	0.169	0.115	0.700	0.077
Боль в груди	0.154	0.147	–	–
Ринит	0.301	0.616	0.375	0.808
Головная боль	0.181	0.012	0.684	–
Диарея	0.111	0.135	0.166	0.443
Боли в животе	0.091	0.082	0.165	0.819
Число заболеваний	5957	1186	1999	525

Таблица 3: Вероятности наличия симптомов для каждого вируса

4 Анализ симптомов

4.1 Точный тест Фишера

В дополнение к проявлению симптомов, сезон рассматривается как дополнительный признак. Для оценки статистической значимости признаков проведен анализ таблиц сопряженности.

Так как проводится анализ наличия зависимости между типом вируса и симптомом для небольших выборок, воспользуемся точным тестом Фишера. Это тест статистической значимости, используемый в анализе категориальных данных, когда размеры выборки малы (являются маленькими). Точный критерий Фишера позволяет проверить гипотезу о независимости двух переменных на основе таблиц сопряженности размером 2 на 2. Выполняется подсчет двустороннего теста с помощью средств языка Python. Полученные результаты вычислений представлены в таблице 4.

Если значение точного критерия Фишера больше критического, принимается нулевая гипотеза и делается вывод об отсутствии статистически значимых различий частоты наличия симптома в зависимости от типа вируса. Если значение точного критерия Фишера меньше критического, принимается альтернативная гипотеза и делается вывод о наличии данной зависимости (12).

Для данного исследования в соответсвии с рекомендациями Министерства Здравоохранения был установлен критический уровень значимости $p = 0.001$, все результаты, которые являются статистически значимыми на этом уровне, выделены в таблице 4 полужирным шрифтом.

	Грипп | РС	Грипп | Адено	Грипп | Бока	РС | Адено	РС | Бока	Адено | Бока
Температура	$\mathbf{3.7 * 10^{-4}}$	$\mathbf{1.4 * 10^{-33}}$	$\mathbf{7.8 * 10^{-25}}$	$\mathbf{3.5 * 10^{-31}}$	$\mathbf{8.2 * 10^{-10}}$	$\mathbf{6.8 * 10^{-62}}$
Кашель	$\mathbf{1.6 * 10^{-33}}$	$\mathbf{6.7 * 10^{-7}}$	$\mathbf{6.0 * 10^{-21}}$	$\mathbf{1.4 * 10^{-13}}$	$9.9 * 10^{-2}$	$\mathbf{1.7 * 10^{-11}}$
Одышка	$\mathbf{2.8 * 10^{-88}}$	$1.6 * 10^{-1}$	$\mathbf{9.1 * 10^{-30}}$	$\mathbf{4.5 * 10^{-48}}$	$1.4 * 10^{-1}$	$\mathbf{5.1 * 10^{-24}}$
Боли в горле	$5.4 * 10^{-2}$	**0.0**	$7.1 * 10^{-1}$	$\mathbf{1.3 * 10^{-132}}$	1.0	$\mathbf{1.6 * 10^{-9}}$
Боли в груди	$4.3 * 10^{-1}$	–	–	–	–	–
Ринит	$\mathbf{1.2 * 10^{-135}}$	$\mathbf{6.3 * 10^{-137}}$	$\mathbf{3.2 * 10^{-38}}$	$9.9 * 10^{-2}$	$\mathbf{8.4 * 10^{-5}}$	$\mathbf{2.2 * 10^{-6}}$
Головная боль	$\mathbf{6.9 * 10^{-7}}$	$\mathbf{1.3 * 10^{-59}}$	–	$\mathbf{1.4 * 10^{-45}}$	–	–
Диарея	$\mathbf{1.7 * 10^{-12}}$	$\mathbf{2.5 * 10^{-45}}$	$\mathbf{2.6 * 10^{-34}}$	$6.1 * 10^{-1}$	$\mathbf{3.6 * 10^{-6}}$	$\mathbf{1.2 * 10^{-7}}$
Боли в животе	$1.5 * 10^{-2}$	$\mathbf{7.9 * 10^{-37}}$	$\mathbf{2.2 * 10^{-64}}$	$4.3 * 10^{-3}$	$\mathbf{6.5 * 10^{-30}}$	$\mathbf{2.2 * 10^{-31}}$
Зима	$\mathbf{9.3 * 10^{-27}}$	$\mathbf{1.6 * 10^{-46}}$	$\mathbf{7.1 * 10^{-16}}$	$\mathbf{1.9 * 10^{-9}}$	$3.9 * 10^{-2}$	$1.4 * 10^{-2}$
Весна	$\mathbf{6.7 * 10^{-18}}$	$9.7 * 10^{-3}$	$\mathbf{4.6 * 10^{-9}}$	$\mathbf{5.7 * 10^{-14}}$	$\mathbf{1.2 * 10^{-21}}$	$\mathbf{8.5 * 10^{-4}}$
Лето	$1.5 * 10^{-3}$	$\mathbf{2.0 * 10^{-76}}$	$\mathbf{8.5 * 10^{-12}}$	$\mathbf{4.6 * 10^{-37}}$	$\mathbf{4.5 * 10^{-5}}$	$\mathbf{7.2 * 10^{-8}}$
Осень	$\mathbf{3.8 * 10^{-10}}$	$\mathbf{1.2 * 10^{-85}}$	$\mathbf{2.6 * 10^{-96}}$	$\mathbf{4.9 * 10^{-33}}$	$\mathbf{1.1 * 10^{-45}}$	$\mathbf{3.3 * 10^{-4}}$

Таблица 4: Точный критерий Фишера, полученный по таблице сопряженности для каждых двух вирусов и каждых симптомов

В результате проведенного исследования установлено, что сезон является значимым дополнительным признаком в анализе заболеваний. Кроме того, статистически значимыми на уровне значимости $p = 0.001$ оказались все симптомы, за исключением "Боль в груди".

5 Наивный байесовский классификатор

Для произведения оценки принадлежности объекта к конкретному типу вируса на основании набора признаков применен алгоритм наивного байесовского классификатора в предположении, что все признаки являются независимыми друг от друга.

При реализации алгоритма необходимо по набору симптомов вычислять относительные вероятности принадлежности к одному из четырех типов вируса. Ранее было получено, что прослеживается зависимость между сезоном и типом вируса. Поэтому сезон будет рассматриваться как отдельный признак, дополняющий таблицу 3. Напомним также, что симптом "Боль в груди" не рассматривается. В результате получается таблица 5. При этом признак "Без учета сезона" оставлен в таблице на случай пропусков в данных.

Тип вируса / Симптом	Грипп	РС	Адено	Бока
Температура	0.789	0.716	0.892	0.615
Кашель	0.744	0.874	0.844	0.918
Одышка	0.400	0.601	0.228	0.557
Боли в горле	0.169	0.115	0.700	0.077
Ринит	0.301	0.616	0.375	0.808
Головная боль	0.181	0.012	0.684	–
Диарея	0.111	0.135	0.166	0.443
Боли в животе	0.091	0.082	0.165	0.819
Произведения вероятностей симптомов	$2.184 * 10^{-5}$	$3.539 * 10^{-6}$	$8.441 * 10^{-4}$	$7.10 * 10^{-3}$
Зима	0.745	0.168	0.052	0.037
Весна	0.643	0.272	0.064	0.021
Лето	0.075	0.092	0.717	0.117
Осень	0.142	0.151	0.382	0.326
Без учета сезона	0.675	0.195	0.083	0.047

Таблица 5: Признаки и вероятности их наличия для каждого из типов вируса

В данной таблице 5 у бокавируса пропущен один признак ("Головная боль"), в связи с этим подсчет вероятностей будет выполняться по схеме, описанной далее.

Разделим типы вирусов на две группы:

- группу 1 (G_1) из вирусов типа Грипп, РС, Адено, для которой известны все признаки
- группу 2 (G_2) из вируса типа Бока, для которой неизвестен признак X_i.

Используем формулу Байеса для наивного байесовского классификатора:

$$P(K_l|x) = \frac{P(K_l)\prod_{i=1}^{n} p_l(x_i)}{\sum_{j=1}^{L} P(K_j)\prod_{i=1}^{n} p_j(x_i)} \qquad (2),$$

где K - тип вируса, $x \in \{0,1\}^8$ - набор признаков, $P(K_l)$ - априорная вероятность наличия вируса K_l в текущий сезон, $p_j(x_i)$ - вероятность наличия (или отсутствия) симптома x_i у вируса j, L - число рассматриваемых вирусов, n - число рассматриваемых симптомов.

Шаг 1

С использованием наивного байесовского классификатора с исключением признака X_i вычисляем относительные вероятности $P(G_1)$ и $P(G_2)$:

$$P(G_1) + P(G_2) = 1$$

Шаг 2

С использованием наивного байесовского классификатора с использованием всех признаков вычисляем относительные вероятности P(Грипп), P(РС) и P(Адено):

$$P(\text{Грипп}) + P(\text{РС}) + P(\text{Адено}) = P(G_1)$$

.

Шаг 3

Относительная вероятность P(Бока):

$$P(\text{Бока}) = P(G_2)$$

.

5.1 Генерация выборки

По таблице 5 сгенерируем выборку. Заметим, что сезон рассматривается в качестве симптома (признака), однако наличие одновременно нескольких сезонов для одного объекта невозможно. Поэтому будем генерировать 10 тысяч объектов для каждого сезона, затем по вероятностям каждого типа вируса для конкретного сезона будет выбран вирус, а далее для конкретного вируса будут сгенерированы 8 симптомов с данными в таблице 5 вероятностями. Затем все объекты в полученной выборке перемешаем случайным образом. В результате получаем таблицу, представленную на рис.4.

	Тип вируса	Сезон	Температура	Кашель	Одышка	Боли в горле	Ринит	Головная боль	Диарея	Боли в животе
0	Грипп	Весна	1	1	1	0	0	0	0	0
1	Бока	Осень	1	1	0	0	1	–	0	1
2	Грипп	Зима	1	1	0	0	1	0	0	0
3	Адено	Осень	1	0	0	1	1	0	0	0
4	Грипп	Весна	0	1	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...
39995	Грипп	Весна	1	0	1	0	0	0	0	0
39996	РС	Зима	1	1	0	0	1	0	0	0
39997	Грипп	Зима	1	1	0	0	0	0	0	0
39998	Грипп	Весна	1	1	1	0	1	0	1	0
39999	Грипп	Осень	0	1	0	0	0	1	0	0

40000 rows × 10 columns

Рис. 4: Полученная выборка

5.2 Качество алгоритма

Рассмотренный ранее позволяет оценить принадлежность объекта к каждому из четырех классов. Для оценки качества работы алгоритма построены 4 ROC-кривых методом "один против всех". Однако, в полученной выборке наблюдается дисбаланс классов, поэтому для каждого типа вируса выбрано такое же число объектов других типов.

Полученные результаты представлены на графиках 5, 6, 7, 8.

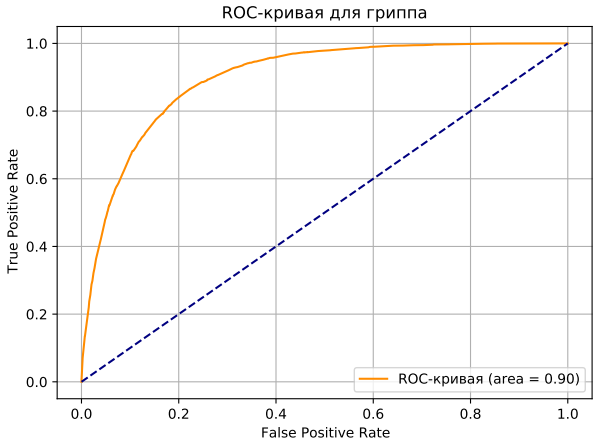

Рис. 5: Кривая ошибок для гриппа

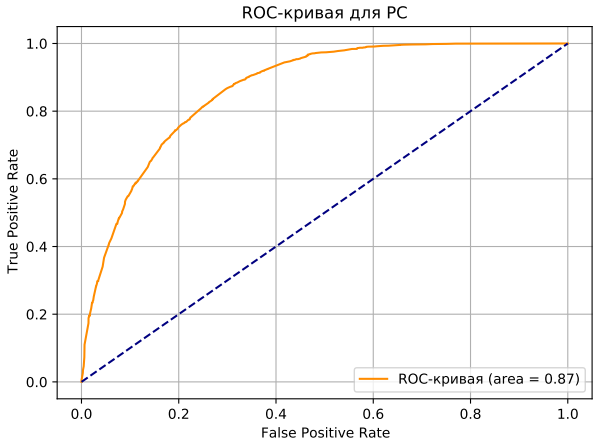

Рис. 6: Кривая ошибок для респираторно-синцитиального вируса

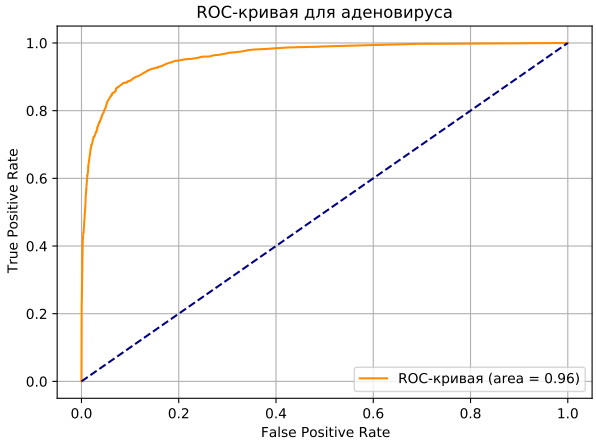

Рис. 7: Кривая ошибок для аденовируса

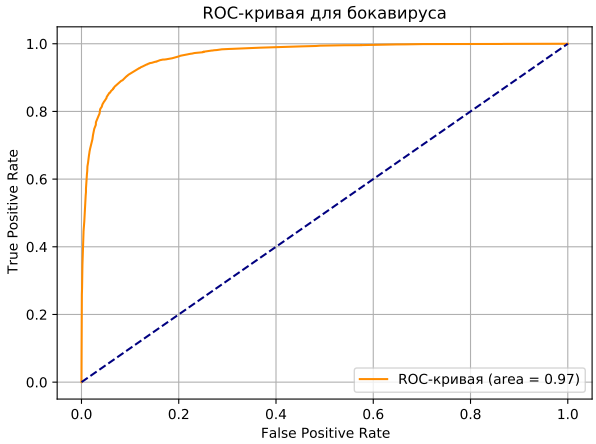

Рис. 8: Кривая ошибок для бокавируса

Список литературы

[1] Материалы ВОЗ

[2] Малый В.П., Романцов М.Г., Сологуб Т.В. Грипп. Пособие для врачей. СПб.-Харьков; 2007.

[3] Сологуб Т.В., Осиновец О.Ю. Иммуномодуляторы в ком- плексной терапии ОРВИ: возможности применения препа- рата галавит. Русский медицинский журнал. 2013;3.

[4] Инфекционные болезни: национальное руководство. Под общ. ред. Ющука Н.Д., Венгерова Ю.Я. М.: Гэотар-Мед.; 2009.

[5] Ершов Ф.И., Касьянова Н.В. Современные принципы про- филактики и лечения гриппа и ОРВИ. Consilium medicum. 2004;1:1-13.

[6] Афанасьева И.А. Гипорамин в лечении ОРВИ у детей. Рус- ский медицинский журнал. 2005;21:1404-1405.

[7] Ершова А.К. Комплексный подход к лечению острых респи- раторных вирусных заболеваний. Русский медицинский жур- нал. 2011;18.

[8] Романцов М.Г., Киселев О.И., Сологуб Т.В. Этиопатогене- тическая фармакотерапия ОРВИ и гриппа. Лечащий врач. 2011;2:92-96.

[9] Трухан Д.И., Мазуров А.Л., Речапова Л.А. Острые респираторные вирусные инфекции: актуальные вопросы диагностики, профилактики и лечения в практике терапевта. Терапевтический архив. 2016;88(11):76-82.

[10] Г. Аптон. Анализ таблиц сопряженности

[11] Р 50.1.033–2001. Рекомендации по стандартизации. Прикладная статистика. Правила проверки согласия опытного распределения с теоретическим. Часть I. Критерии типа хи-квадрат. – М.: Изд-во стандартов. 2002. – 87 с.

[12] Библиотека постов MEDSTATISTIC об анализе медицинских данных. ТОЧНЫЙ КРИТЕРИЙ ФИШЕРА.

[13] Акад. РАН Ю.В. БЕЛОВ, к.м.н. Г.И. САЛАГАЕВ, к.м.н. А.В. ЛЫСЕНКО, к.м.н. П.В. ЛЕДНЕВ. Мета-анализ в медицине. Хирургия 3, 2018.

[14] А. М. Гржибовский. ДОВЕРИТЕЛЬНЫЕ ИНТЕРВАЛЫ ДЛЯ ЧАСТОТ И ДОЛЕЙ. Экология человека 2008.05. УДК 31:61