МОДЕРНИЗАЦИЯ ПРОГНОСТИЧЕСКИХ РЕГРЕССИОННЫХ МОДЕЛЕЙ ДЛЯ ОЦЕНКИ КОЛИЧЕСТВА ЛЕТАЛЬНЫХ ИСХОДОВ ПРИ НОВОЙ КОРОНАВИРУСНОЙ ИНФЕКЦИИ

DOI: 10.17691/stm2020.12.4.01
УДК 578.834.1:616–036.2/.8:303.724.32
Поступила 4.06.2020 г.

© **Н.Н. Карякин,** д.м.н., ректор[1];
Н.В. Саперкин, к.м.н., доцент кафедры эпидемиологии, микробиологии и доказательной медицины[1];
А.П. Баврина, к.б.н., доцент кафедры медицинской физики и информатики[1];
О.В. Другова, к.б.н., доцент кафедры медицинской физики и информатики[1];
В.И. Климко, к.т.н., главный специалист[2];
А.С. Благонравова, д.м.н., проректор по научной работе[1];
О.В. Ковалишена, д.м.н., профессор, зав. кафедрой эпидемиологии, микробиологии и доказательной медицины[1]

[1]Приволжский исследовательский медицинский университет, пл. Минина и Пожарского, 10/1, Н. Новгород, 603005;
[2]ГК «МедИнвестГрупп», ул. Александра Солженицина, 27, Москва, 109004

Цель исследования — модернизировать созданные прогностические регрессионные модели в условиях расширения знаний о новой коронавирусной инфекции COVID-19.

Материалы и методы. В основу модификации моделей и повышения их предсказательной способности положен мониторинг открытых данных из международных и российских информационных баз. Вычислены традиционные описательные статистики, для моделирования использовали линейный регрессионный анализ. Работы выполнены с помощью программ IBM SPSS Statistics 26.0 и R 3.6.0 (RStudio).

Результаты. Изучены проявления эпидемического процесса заболеваемости COVID-19 в нескольких странах с особым вниманием к возникновению случаев летальных исходов, ассоциированных с данной инфекцией. Отмечен значительный процент тяжелых форм среди заболевших как в России, так и за рубежом. С учетом достижения пика заболеваемости в Китае и Италии авторы провели работу над усовершенствованием представленных ранее (см. журнал «Современные технологии в медицине», Т. 12, №2, 2020 г.) регрессионных моделей и сравнили их эффективность. Первая модифицированная модель основана на абсолютном приросте новых случаев инфекции: регрессионный коэффициент равен 0,16 (95% ДИ 0,137–0,181). Эти сведения относятся к базовой информации, которая аккумулируется в открытых источниках. В расширенной версии обновленной модели кроме указанного фактора также учитывали данные о случаях тяжелого течения инфекции: регрессионные коэффициенты 0,128 (95% ДИ 0,103–0,153) и 0,053 (95% ДИ 0,029–0,077) соответственно, p=0,0001 при сравнении модели 2 с моделью 1.1.

Заключение. Основываясь на новых текущих (с января по май 2020 г.) данных о заболеваемости COVID-19 в мире и отдельных странах, авторы выполнили конкретизацию исходной и расширенной регрессионных моделей прогнозирования случаев летальных исходов. Полученные оптимизированные модели экстраполированы на новую ситуацию по инфекции, что позволит и далее совершенствовать наш аналитический подход. В настоящее время продолжается сбор данных для улучшения предсказательной способности моделей.

Ключевые слова: COVID-19; SARS-CoV-2; коронавирус; прогнозирование исхода инфекции; многомерная регрессионная модель; прогнозирование летальности инфекции.

Для контактов: Баврина Анна Петровна, e-mail: annabavr@gmail.com

Как цитировать: Karyakin N.N., Saperkin N.V., Bavrina A.P., Drugova O.V., Klimko V.I., Blagonravova A.S., Kovalishena O.V. Modernization of regression models to predict the number of deaths from the new coronavirus infection. Sovremennye tehnologii v medicine 2020; 12(4): 6–12, https://doi.org/10.17691/stm2020.12.4.01

English

Modernization of Regression Models to Predict the Number of Deaths from the New Coronavirus Infection

N.N. Karyakin, MD, DSc, Rector[1];
N.V. Saperkin, MD, PhD, Associate Professor, Department of Epidemiology, Microbiology and Evidence-Based Medicine[1];
A.P. Bavrina, PhD, Associate Professor, Department of Medical Physics and Informatics[1];
O.V. Drugova, PhD, Associate Professor, Department of Medical Physics and Informatics[1];
V.I. Klimko, PhD, Chief Specialist[2];
A.S. Blagonravova, MD, DSc, Vice-Rector for Science[1];
O.V. Kovalishena, MD, DSc, Professor, Head of Department of Epidemiology, Microbiology and Evidence-Based Medicine[1]

[1]Privolozhsky Research Medical University, 10/1 Minin and Pozharsky Square, Nizhny Novgorod, 603005, Russia;
[2]GC "MedInvestGroup", 27 Alexander Solzhenitsyn St., Moscow, 109004, Russia

The aim of the study was to modernize the existing prognostic regression models in the context of expanding knowledge about the new coronavirus infection.

Materials and Methods. The modification of models and the increase in their predictive ability are based on collecting the available data from international and Russian databases. We calculated the traditional descriptive statistics and used the linear regression analysis for modeling. The work was performed using the IBM SPSS Statistics 26.0 and the R 3.6.0 (RStudio) software.

Results. Manifestations of the COVID-19 epidemic process in several countries were studied; special attention was put to the number of deaths associated with the infection. A significant proportion of severe cases were noted among patients both in Russia and elsewhere. Considering that the disease incidence has reached its peak in China and Italy, we were able to improve the previously published (Sovremennye tehnologii v medicine 2020, Vol. 12, No.2) regression models and to compare their performance. The first modified model is based on the absolute increase in new cases of the infection: its regression coefficient is 0.16 (95% CI 0.137–0.181). In the extended version of the updated model, we additionally considered cases of aggravated COVID-19: the regression coefficients were 0.128 (95% CI 0.103–0.153) for model 2 and 0.053 (95% CI 0.029–0.077) for model 1.1; p=0.0001.

Conclusion. Based on the most recent data (from January to May 2020) on the incidence of COVID-19 in the world, we have developed more specific versions of the basic and extended regression models of lethal outcomes. The resulting models are optimized and extrapolated to the current epidemiological situation; they will allow us to improve our analytical approach. For that purpose, data collection is currently ongoing.

Key words: COVID-19; SARS-CoV-2; coronavirus; predicting the outcome of infection; multivariate regression model; predicting infection-associated mortality.

Введение

В ответ на продолжающееся распространение COVID-19 в разных странах в мире не прекращается интенсивная деятельность по обеспечению адекватного ответа на риски, связанные с влиянием этой инфекции на системы здравоохранения. Эффективным способом решения подобных задач является, как известно, составление краткосрочных и долгосрочных прогнозов развития эпидемиологической ситуации. Процесс формирования заболеваемости по сравнению с начальным периодом эпидемии новой коронавирусной инфекции и наше понимание его закономерностей претерпевают ряд существенных изменений. Прежде всего они связаны с возрастающими диагностическими возможностями, влиянием социально-ограничительных противоэпидемических мероприятий, а также меняющимися подходами к регистрации случаев заболевания и летального исхода [1–6].

Математическое моделирование широко применяется при изучении эпидемиологии COVID-19. Такой подход позволяет получать ответы на ряд первостепенных вопросов, связанных с характеристикой эпидемического процесса в динамике, оценкой эффективности противоэпидемических мероприятий, а также с определением потребностей практического

здравоохранения в силах и средствах диагностики, лечения и профилактики.

В России после первых завозных случаев новой коронавирусной инфекции в январе 2020 г. количество выявленных заболеваний стало превышать 20 случаев в начале марта и к маю все субъекты страны уже столкнулись с COVID-19. Первые летальные исходы, вызванные этим заболеванием, стали фиксировать с конца марта. Сейчас случаи инфекции регистрируются в городах и сельской местности, среди различных возрастных групп. Важно отметить вспышки инфекции, возникшие в медицинских организациях, а также среди людей, работающих вахтовым методом [7, 8]. Поэтому в условиях постоянно изменяющихся данных, особенно после выхода некоторых стран на плато, и с началом спада числа выявления новых случаев COVID-19 предложенные ранее математические модели [9] потребовали существенной модификации.

Цель исследования — модернизировать разработанные ранее регрессионные модели для прогнозирования летальных исходов в условиях дальнейшего развития эпидемического процесса.

Материалы и методы

В представленном эпидемиологическом исследовании взяты данные из открытых источников, которые размещены на соответствующих официальных сайтах в сети Интернет. Детали получения необходимой количественной информации описаны в предыдущей публикации [9].

Статистическую обработку данных проводили с помощью лицензионного программного обеспечения IBM SPSS Statistics 26.0 и R 3.6.0 (RStudio) (пакет RVAideMemoire). Проверку нормальности распределения осуществляли с помощью критерия Колмогорова–Смирнова и построения квартильных диаграмм (графика квартилей — Q–Q-plot). Силу связи оценивали с помощью коэффициента корреляции Спирмена, характер связи — с помощью простой и множественной линейной регрессии. Результаты представлены в виде M±SD, где M — среднее, SD — стандартное отклонение; Me [МКИ], где Me — медиана, МКИ — межквартильный интервал (Q1–Q3), и в виде абсолютных значений в арифметической и логарифмической шкалах; процентные доли представляли с указанием стандартного отклонения процентной доли ($P\pm\sigma_p\%$). За критический уровень значимости принят $p\leq0,05$. При необходимости рассчитывали 95% доверительный интервал (ДИ). Сравнение моделей проводили с помощью дисперсионного анализа ANOVA и информационного критерия Акаике (AIC).

Результаты и обсуждение

В мире зарегистрировано уже свыше 5 млн. заболевших COVID-19 (по данным на 30 мая 2020 г.), в том числе более 362 тыс. смертей [2]. В России на эту же дату насчитывается 405 843 лабораторно подтвержденных случаев инфекции (с максимальными значениями в Москве — 180 791, при этом в Нижегородской области выявлено 9834 случая). На долю летальных исходов в России приходится 1,16±0,02% от всех подтвержденных случаев [8]. При анализе абсолютного прироста случаев COVID-19 (в период с 30 января до 1 июня 2020 г.) отмечается начало постепенного роста числа выявленных заболеваний в России с начала апреля, т.е. позже, чем в Италии и США (рис. 1). Кроме того, обращают на себя внимание существенные различия места изучаемых стран по уровню зарегистрированной заболеваемости: показатели для США значительно превышают таковые для остальных стран на протяжении всего периода наблюдения. Необходимо отметить, что в Китае начиная со второй половины февраля отмечено отчетливое снижение вновь выявляемых случаев заболевания этой инфекцией. Максимальный прирост заболеваемости в абсолютном выражении (3893 случая) в Китае зафиксирован 13 февраля 2020 г.

При изучении характера появления новых летальных случаев выявлены различия между странами по времени начала регистрации первых смертей от коронавирусной инфекции, амплитуде показателей, а также моменту снижения абсолютного прироста летальных исходов (рис. 2). Вслед за Китаем достаточно рано первые летальные исходы стали фиксироваться в Италии (23 февраля 2020 г.) и США (3 марта 2020 г.). В связи с поздним началом эпидемического процесса COVID-19 в России первые смертельные исходы стали появляться только с конца марта.

Если в первой трети пандемии летальные исходы возникали лишь в Китае и находились ежедневно на уровне не более 260 случаев, то с середины марта стала лидировать Италия, где абсолютные числа превышали 900 случаев. Необходимо отметить, что уже с начала апреля и по сегодняшний день максимальные значения летальных исходов наблюдаются в США и значительно превышают другие три страны. В США в настоящее время отмечаются резкие колебания по числу смертей, при этом абсолютный прирост стабильно превышает 1000 случаев за исключением отдельных дней наблюдения.

Ниже приведено статистическое описание последних данных (заболеваемость и летальность) по COVID-19 в Китае и Италии в период после достижения пика по количеству выявленных случаев новой коронавирусной инфекции (табл. 1 и 2). Выбор указанных территорий обусловлен отсутствием спада заболеваемости в России и США на момент проведения исследования.

С учетом характера распределения, отличающегося от нормального, для описания вариационного ряда использована медиана: в анализируемый период (после пика) медианное значение случаев коронавирусной инфекции в Италии превышает таковое для

данных по Китаю более, чем в 14 раз (рис. 3). Обращает на себя внимание значительная разница в зарегистрированных новых случаях между Италией и Китаем.

Вариационный ряд значений абсолютного прироста летальных исходов был проверен на соответствие нормальному распределению. Результаты проверки представлены на рис. 4. При анализе графиков отчетливо видно различие данных по двум странам в плане характера распределения количества летальных исходов: данные официальной регистрации в США подчиняются закону нормального распределения, а результаты по Китаю распределены несимметрично. Это может быть связано с изменением тяжести инфекции в Китае, что сказывается в прекращении появления летальных исходов при наличии COVID-19.

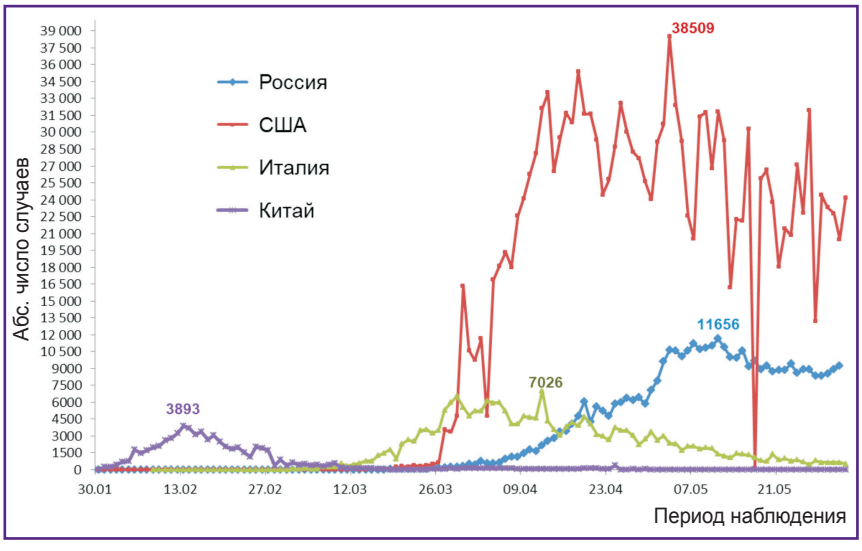

Рис. 1. Динамика выявления новых случаев COVID-19 с января по май 2020 г. в разных странах (по официальным данным)

После прохождения пика по заболеваемости абсолютный прирост летальных исходов в Китае характеризуется медианным значением, равным 2 случаям [МКИ 0–14], в то время как в Италии ситуация существенно отличается и в среднем в этой стране регистрируют 346,1±261,6 случаев смертей,

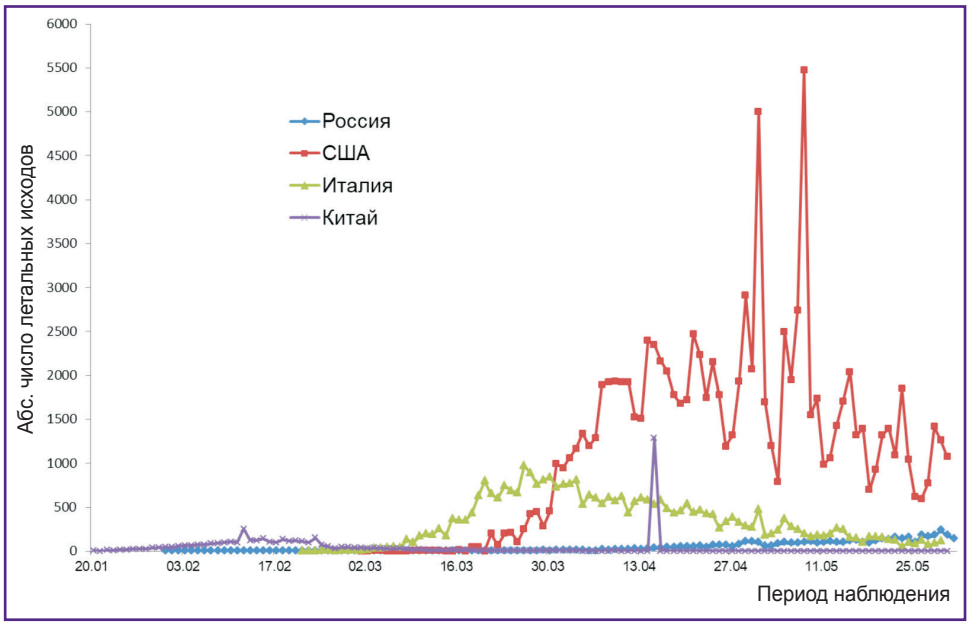

Рис. 2. Абсолютный прирост летальных случаев, обусловленных COVID-19, с января по май 2020 г. в разных странах (по официальным данным)

Таблица 1

Описательные статистики для вновь выявленных случаев COVID-19 после достижения пика заболеваемости

Страна	Me [МКИ]	Минимум	Максимум	95% ДИ
Китай	117 [45–364]	11	894	133–262
Италия	1739 [882–2698]	451	4092	882–2698

Таблица 2

Описательные статистики для абсолютного прироста летальных исходов

Китай	Me [МКИ]	Минимум	Максимум	95% ДИ
	2 [0–14]	0	150	9–22
Италия	M±SD	Минимум	Максимум	95% ДИ
	346,1±261,6	0	971	292–400

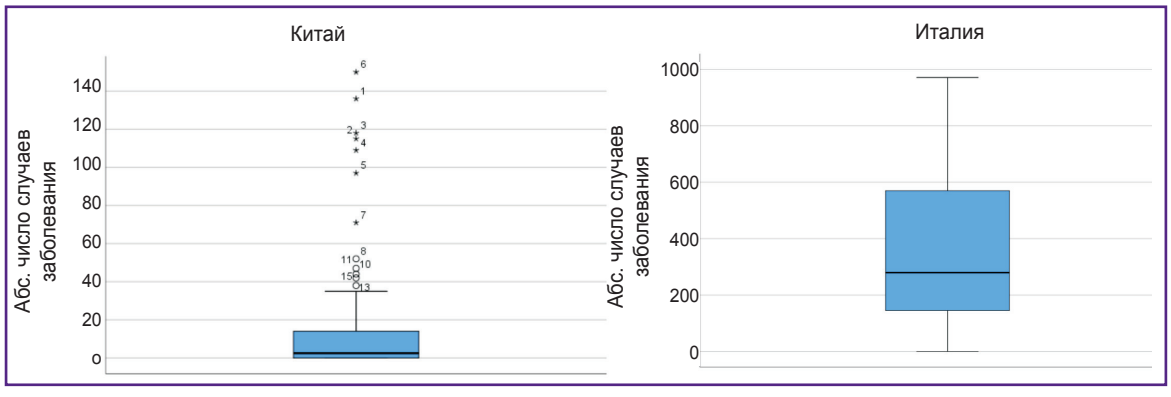

Рис. 3. Характеристика распределения значений абсолютного прироста вновь зарегистрированных случаев COVID-19

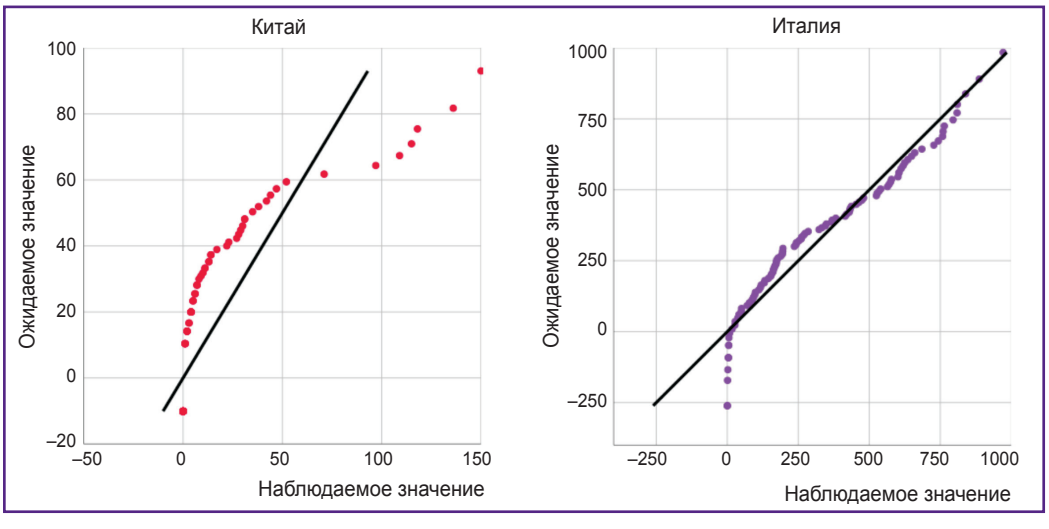

Рис. 4. Проверка нормальности распределения с помощью графиков квартилей (Q–Q-plot)

Таблица 3

Характеристики новой модели 1.1

Константа	Значение константы	Стандартная ошибка	Уровень значимости	95% ДИ
β_0	–1,285	2,172	0,556	–5,603...3,034
β_1	0,16	0,011	0,0001	0,137...0,181

Таблица 4

Характеристики новой модели 1.2

Константа	Значение константы	Стандартная ошибка	Уровень значимости	95% ДИ
β_0	–0,873	0,579	0,137	–2,008...0,262
β_1	0,749	0,124	0,0001	0,505...0,992

обусловленных COVID-19. В Китае количество смертей не превышало 150 случаев, а в Италии оно приближается к тысяче.

Предыдущее исследование [9] показало наличие сильной корреляции между появлением новых случаев коронавирусной инфекции, зарегистрированных в разных странах. Это позволило выполнить разработку прогностических моделей. Нами были построены эффективные модели, позволяющие проводить расчет количества летальных исходов при новой коронавирусной инфекции при приближении рассматриваемых стран к пику эпидемии. Однако выход показателей инцидентности на плато и начало спада выявления новых случаев COVID-19 потребовали обновления разработанных моделей с учетом текущих данных.

Новой одномерной модели (модель 1.1), основанной на аналогичной исходной модели [9], соответствует уравнение $Y=X\cdot0,16–1,285$ с коэффициентом детерминации $R^2=0,686$ (табл. 3) и AIC=809,37.

Следующий вариант этой регрессионной модели (1.2) предусматривает проведение логарифмического преобразования, что дает возможность соблюсти допущение о наличии линейной ассоциации между переменными (табл. 4).

После подстановки коэффициентов уравнение линейной регрессии (модель 1.2) приобрело вид: $\ln(Y)=\ln(X)\cdot 0{,}749–0{,}873$.

С учетом особенностей логарифмирования при нулевых значениях коэффициент детерминации R^2 составил 0,400, критерий AIC — 186,45.

Необходимо отметить, что приведенные выше модели (1.1 и 1.2) можно использовать только при небольшом количестве вновь выявляемых случаев инфекции (не более 30).

Модернизация разработанных ранее моделей подразумевала их расширение за счет включения дополнительных независимых переменных. Как было показано в предыдущей публикации [9], добавление дополнительной информации — ежедневного абсолютного прироста тяжелых форм COVID-19 — приводит к увеличению точности модели с сохранением ее экономичности. С учетом обновленных расчетов были получены следующие характеристики модели (табл. 5).

Новая модель 2 примет следующий вид: $Y=X_1\cdot 0{,}057+X_2\cdot 0{,}04–9{,}76$.

Также мы попытались расширить модель 2 с помощью видоизмененной версии (квадратная трансформация) указанных переменных, но это не привело к существенному увеличению ее эффективности.

В целом дополнение модели информацией о количестве тяжелых форм не только привело к увеличению коэффициента детерминации R^2 до 0,741, но и повысило точность модели (AIC=793,3). Результаты сравнения моделей с помощью ANOVA указывают на то, что более сложная модель 2 статистически значимо лучше описывает реальную ситуацию, чем однофакторная модель (F-статистика=19,285; p=0,0001).

Ниже приведены результаты тестирования новой модели 2 на предмет выполнения условия присутствия линейной ассоциации между независимой и зависимой переменными. Соответствие этому допущению представлено на рис. 5, демонстрирующем характер распределения нестандартизованных остатков. Отмечается равномерное распределение числовых данных, при котором дисперсия остатков существенно не меняется с увеличением предсказываемой величины, а следовательно, условие линейности ассоциаций для регрессионной модели выполняется.

Была произведена проверка модели на количестве летальных случаев, которые зафиксированы в Китае в определенные даты. Она показала следующие результаты.

Пример 1. После прохождения пика эпидемии в Китае по состоянию на 28.02.2020 г. имеем:

$Y=0{,}128\cdot 435+0{,}053\cdot 288–5{,}857=65$ (хотя на эту дату наблюдалось 44 случая, порядок числовой величины сохраняется).

Пример 2. По состоянию на 5.03.2020 г.:

$Y=0{,}128\cdot 29+0{,}053\cdot 194–5{,}857=8$ (на эту дату отмечено 10 летальных случаев, через четыре дня было выявлено 8 случаев).

Пример 3. По состоянию на 09.03.2020 г.:

$Y=0{,}128\cdot 20+0{,}053\cdot 317–5{,}857=13$ (на эту дату зафиксировано 23 случая, но спустя 4 дня, т.е. 12.03.2020 г., количество летальных исходов — 13).

Важно обратить внимание на присутствие определенного временно́го сдвига в появлении интересуемого исхода, что планируется принять к сведению при дальнейшей актуализации моделей.

Таблица 5

Характеристики новой модели 2

Константа	Значение константы	Стандартная ошибка	Уровень значимости	95% ДИ
β_0	–5,857	2,2	0,01	–10,311...–1,402
β_1	0,128	0,012	0,0001	0,103...0,153
β_2	0,053	0,013	0,0001	0,029...0,077

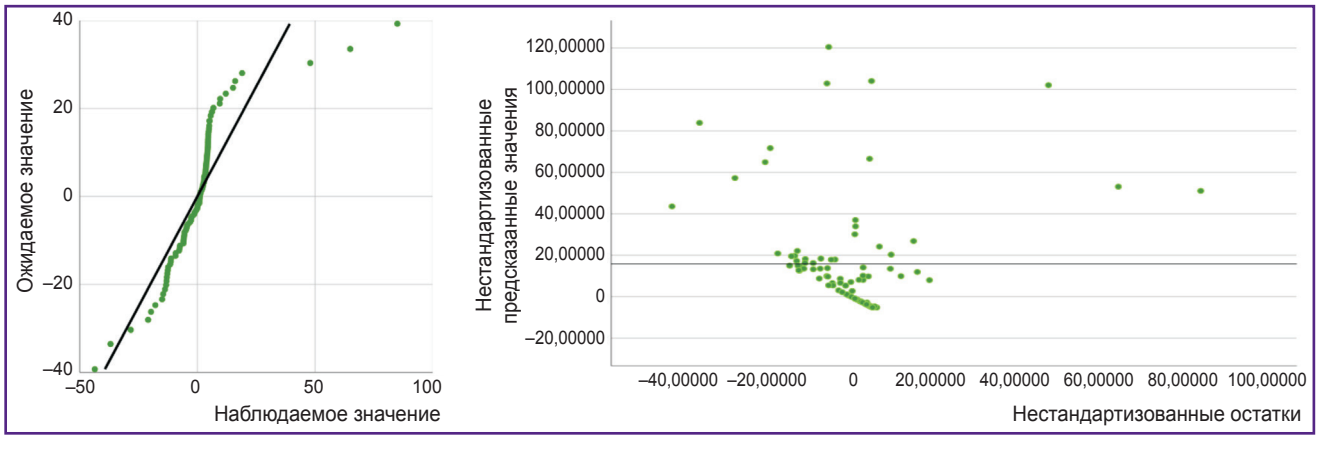

Рис. 5. График квартилей (Q–Q-plot) для регрессионных остатков и распределение прогнозируемых значений в зависимости от остатков (для модели 2)

Заключение

В настоящее время в разных странах мира разработано несколько десятков математических моделей с целью прогнозирования тенденций эпидемического процесса COVID-19, определения эффективности противоэпидемических и профилактических мероприятий и оценки потребностей системы здравоохранения. С учетом новых текущих данных о заболеваемости COVID-19 в мире и отдельных странах была проведена модернизация разработанных (исходной и расширенной) регрессионных моделей прогнозирования случаев летальных исходов. С этой целью был детально изучен период пандемии в Италии и Китае после достижения этими государствами пиковых показателей заболеваемости. В результате были обновлены регрессионные коэффициенты для интерсепта и выбранных независимых переменных. В настоящее время продолжается дальнейший сбор данных для улучшения предсказательной способности моделей, основанных на использовании информации из открытых источников.

Финансирование исследования и конфликт интересов. Исследование не финансировалось каким-либо источником, и конфликты интересов, связанные с данным исследованием, отсутствуют.

Литература/References

1. WHO/HQ/DDI/DNA/CAT. *International guidelines for certification and classification (coding) of COVID-19 as cause of death. Based on International Statistical Classification of Diseases (16 April 2020).* URL: https://www.who.int/classifications/icd/Guidelines_Cause_of_Death_COVID-19.pdf?ua=1.

2. World Health Organization. *Coronavirus disease (COVID-19). Situation report — 131*. URL: https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200530-covid-19-sitrep-131.pdf?sfvrsn=d31ba4b3_2.

3. Centers for Disease Control and Prevention. *Interim guidelines for collecting, handling, and testing clinical specimens for COVID-19.* URL: https://www.cdc.gov/coronavirus/2019-nCoV/lab/guidelines-clinical-specimens.html.

4. *СП 3.1.3597-20 «Профилактика новой коронавирусной инфекции (COVID-19)».*

SP 3.1.3597-20 "Profilaktika novoy koronavirusnoy infektsii (COVID-19)" [SP 3.1.3597-20 "Prevention of a new coronavirus infection (COVID-19)"].

5. *Методические рекомендации МР3.1.0178-20 «Определение комплекса мероприятий, а также показателей, являющихся основанием для поэтапного снятия ограничительных мероприятий в условиях эпидемического распространения COVID-19».*

Metodicheskie rekomendatsii MR3.1.0178-20 "Opredelenie kompleksa meropriyatiy, a takzhe pokazateley, yavlyayushchikhsya osnovaniem dlya poetapnogo snyatiya ogranichitel'nykh meropriyatiy v usloviyakh epidemicheskogo rasprostraneniya COVID-19" [Methodical recommendations MP3.1.0178-20 "Definition of a set of measures, as well as indicators that are the basis for the phased removal of restrictive measures in the context of the epidemic spread of COVID-19"].

6. *Письмо Управления Роспотребнадзора по Нижегородской области №52-00-08/03-3329-2020 от 27.04.2020 г. «Об учете коронавирусной инфекции».*

Pis'mo Upravleniya Rospotrebnadzora po Nizhegorodskoy oblasti No.52-00-08/03-3329-2020 ot 27.04.2020 g. "Ob uchete koronavirusnoy infektsii" [Letter of the Rospotrebnadzor in the Nizhny Novgorod Region No.52-00-08/03-3329-2020 dated on April 27, 2020 "On accounting for coronavirus infection"].

7. Министерство здравоохранения Российской Федерации. *Профилактика, диагностика и лечение новой коронавирусной инфекции (COVID-19). Временные методические рекомендации. Версия 6 (28.04.2020).* URL: https://static-1.rosminzdrav.ru/system/attachments/attaches/000/050/116/original/28042020_%D0%9CR_COVID-19_v6.pdf.

Ministry of Health of the Russian Federation. *Profilaktika, diagnostika i lechenie novoy koronavirusnoy infektsii (COVID-19). Vremennye metodicheskie rekomendatsii. Versiya 6 (28.04.2020)* [Prevention, diagnosis and treatment of new coronavirus infection (COVID-19). Temporary guidelines. Version 6 (April 28, 2020)]. URL: https://static-1.rosminzdrav.ru/system/attachments/attaches/000/050/116/original/28042020_%D0%9CR_COVID-19_v6.pdf.

8. Министерство здравоохранения Российской Федерации. *Информационный ресурс о COVID-19.* URL: https://covid19.rosminzdrav.ru/.

Ministry of Health of the Russian Federation. *Informatsionnyy resurs o COVID-19* [Information resource about COVID-19]. URL: https://covid19.rosminzdrav.ru/.

9. Melik-Huseynov D.V., Karyakin N.N., Blagonravova A.S., Klimko V.I., Bavrina A.P., Drugova O.V., Saperkin N.V., Kovalishena O.V. Regression models predicting the number of deaths from the new coronavirus infection. *Sovremennye tehnologii v medicine* 2020; 12(2): 6–13, https://doi.org/10.17691/stm2020.12.2.01.