О СВЯЗИ ОНКОЛОГИЧЕСКОЙ ЗАБОЛЕВАЕМОСТИ В РЕГИОНАХ СО СТАРЕНИЕМ НАСЕЛЕНИЯ, ЗАГРЯЗНЕНИЯМИ ВОЗДУХА И ВОДЫ

Степанов В.С. (Москва)ⁱ

Борьба с ростом злокачественных новообразований (ЗНО) является в России и многих других странах весьма актуальной задачей. Заболеваемость - контингент страдающих ЗНО - увеличилась с 2005 года за полтора десятка лет в 1.5 раза: от 1680 до 2677 чел. в расчёте на 10^5 россиян (МНИОИ, 2020). По определению, контингент Y включает “численность лиц со всеми локализациями ЗНО, которые состоят на учёте в лечебных организациях Минздрава” в году t (ЗД-РФ, 2019). Рост заболеваемости Y во многих регионах связан с неблагоприятной экологией, со старением населения и рядом показателей: по особенностям диеты граждан, образа их жизни и др. (Бузинов, 2016), (Сулейманов, 2019). Ещё одной особенностью таких заболеваний является их многофакторная природа. Поэтому для изучения факторов риска по различным базам онкологических данных в последнее время стали применять методы многомерной статистики, которые реализованы в программных системах *IBM SPSS*, *Statistica* и прочих (Мурасова, 2017), (Лазарев, 2017), (Айдинов, 2015), (Фельдблюм, 2018).

Целью работы было построение (с оценкой по точности) модели линейной регрессии, которая связывает региональный уровень заболеваемости Y^t к концу года t с комплексом факторов, наблюдаемых в регионе с лагами относительно отсчёта t . При выборе лагов учитывалось, что каждый фактор риска способствует появлению ЗНО не синхронно, а с некоторым запаздыванием (Сулейманов, 2019). Подобная модель может быть полезной при оценке экологических платежей для предприятий, т.к. связывает онкологическую заболеваемость с объёмами годовых сбросов загрязненной воды в природную среду региона, с загрязнениями воздуха в населенных пунктах.

С учетом сказанного, факторы (факторные переменные) для модели можно частично сразу взять или легко рассчитать из региональных таблиц Росстата:

1. X_1 – доля пожилых (лица старше трудоспособного возраста, в [%]) (РР-СЭП, 2020);
2. X_2 – доля исследованных проб воздуха в населенных пунктах, в которых превышена предельно допустимая концентрация (ПДК) вредных веществ, [%] (эта доля есть характеристика гигиенического состояния атмосферы в регионе) (ООС-РФ, 2018);
3. $X_3 = V_{зсв}/S$ – удельный сброс загрязненных сточных вод (ООС-РФ, 2018), где $V_{зсв}$ – объём сброса этих вод в регионе в году t (в [млн. м³]; берётся взвешенная сумма вод “без очистки” и “очищенных ниже норматива”), S – площадь территории региона в [тыс. км²], за вычетом площади его лесов, дорог и населенных пунктов (РР-СЭП, 2020).

В материал обучения – выборку из информационной базы, – на основе которого оценивались параметры модели, зависимая переменная Y входила за 2017 и 2018 годы (ЗД-РФ, 2019). Затем, на этапе “экзамен” модель применялась для оценки Y в 2019 году.

При изучении взаимосвязи с заболеваемостью 4-го фактора (X_4) – по обеспечению граждан безопасной питьевой водой – оказалось, что регионы неоднородны. Например, в группу 1, включающую две столицы РФ, попали также Липецкая, Орловская, Пензенская, Рязанская области, пострадавшие после аварии на ЧАЭС (Словарь, 2003), и, Алтайский край – после давних семипалатинских взрывов; ещё – республики Коми, Татарстан, Мордовия; Камчатский край и ряд областей: Воронежская, Ивановская, Кемеровская, Курская, Мурманская, Оренбургская, Свердловская, Ульяновская, Ярославская. В них хорошо развита обрабатывающая и добывающая промышленность (либо имеется повышенный радиоактивный фон). Поэтому в этих регионах, несмотря на довольно высокую безопасность питьевой воды, наблюдаются повышенные значения Y . На рис.1 показаны точки (X_4, Y) для этих регионов, к которым был подогнан линейный тренд.

Рис.1. Зависимость Y^t в 2019 г. от X_4 в 2017 г., в [%]

В группе 2 из 26-ти регионов, значения X_4 почти всегда между 80% и 95%, а Y ниже. Из них 7 регионов – из ЦФО, 6 регионов из ПФО и ещё здесь Красноярский, Приморский края; Челябинская, Новосибирская, Омская, Томская, Иркутская, Волгоградская, Ростовская, Магаданская области. Линейный тренд $Y^t = -74,193 \cdot X_4^{t-2} + 9438$ имел $R^2 = 87.6\%$. В группе 3 из 7-ми регионов, фактор X_4 лежит между 78% и 92%. Сюда вошли республики Башкортостан, Крым, Карелия и Сахалинская, Смоленская, Тверская и Тюменская (без АО) области. Тренд здесь был $Y^t = -70,605 \cdot X_4^{t-2} + 8710$ с $R^2 = 97.9\%$, т.е. тренды для 2-й и 3-й групп почти параллельны друг другу. Все эти тренды показывают, что с ухудшением гигиены питьевой воды обычно растёт заболеваемость Y^t .

Переменные $X_1, \dots, X_4$ оказались значимыми в “модели взаимодействий факторов окружающей среды” при анализе распространения новообразований (Бузинов, 2016). Так, здесь находились корреляции между заболеваемостью ЗНО в населенных пунктах Примо-

рья и каждым из десятков показателей среды обитания. Затем на основе теории информации установлено, что сильнее всего на появление ЗНО повлияли: демографическая структура (фактор X_1); качество питьевой воды (X_4); объёмы сброса загрязненной воды (X_3), загрязнение воздуха (X_3). Пара X_1, X_4 входила в информативные в (Айдинов, 2015), (Лазарев, 2017), (Мирасова, 2017), (Фельдблюм, 2018).

Модель оценивалась не по всему набору регионов из информационной базы, собранной нами с 2012 по 2019 гг. В частности, сначала были исключены многие из них. В обучающую таблицу вошло 65 строк с наборами переменных, относящихся к республикам: Крым и Татарстан; 7-ми краям: Алтайский, Забайкальский, Краснодарский, Красноярский, Пермский, Приморский, Хабаровский, ХМАО, многим областям. Области часто были в составе ЦФО, реже в СЗФО, ПФО, ЮФО и совсем редко в СФО, ДФО (Амурская, Архангельская, Белгородская, Брянская, Волгоградская, Воронежская, Ивановская, Иркутская, Костромская, Курганская, Курская, Липецкая, Московская, Нижегородская, Новгородская, Омская, Орловская, Пензенская, Псковская, Ростовская, Рязанская, Самарская, Сахалинская, Тверская, Томская, Тюменская, Ярославская). Примерно $\frac{3}{4}$ регионов материала обучения включает более 90% этноса “русские”; кроме того, общие коэффициенты смертности во многих исключенных регионах были ниже 28-го места в 2018 году; в исключенных регионах часто проживает много молодежи (PP-ОХ, 2019), (PP-СЭП, 2020).

В итоге оценивалась линейная модель $Y = \beta_0 + \beta_1 \cdot X_1 + \dots + \beta_3 \cdot X_3 + \varepsilon$, где Y – уровень заболеваемости ЗНО в регионе на конец года t ; $X_1, \dots, X_3$ – тройка вышеописанных “факторов”; переменная ε отражает влияние на Y неизвестных нам факторов, а $\beta_0, \dots, \beta_3$ – числовые параметры. Величина ε определяется как случайная со средним 0 и дисперсией d , которая неизвестна. Оценки всех параметров: $\beta_0, \dots, \beta_3, d$, – находятся по обучающей таблице посредством МНК, который был реализован функциями *Excel* (Карлберг, 2018). Обозначим b_j – МНК- оценки β_j , что находятся по обучающей таблице. Каждая её строка содержит наборы вида $Y, X_1, \dots, X_3$ для регионов, причём значения X_j входят с лагом относительно отсчёта t : от 1-го до 3-х лет для разных X_j . Переменная Y имеет сверху индекс, привязанный к метке года t ; выше он для простоты записи не показывался. Всего в состав таблицы с обучающими данными вошло $n = 65$ строк; немного менее $\frac{1}{2}$ были для Y за 2017 год, а прочие были – для Y за 2018 г.

Исходя из вышеописанного, выражение модели с МНК- оценками и явным обозначением лагов, имеет вид

$$Y^t = b_0 + b_1 \cdot X_1^{t-1} + b_2 \cdot X_2^{t-2} + b_3 \cdot X_3^{t-3}, \quad (1)$$

где Y^t (или Y) – контингент больных со ЗНО к концу года t в некотором регионе РФ, который имеет значения факторов $X_1, \dots, X_3$, взятых с лагом относительно t , а b_j – оценки параметров β_j (их численные значения приводятся в строке 1 из Таб.).

Таблица Результаты оценивания параметров и статистики

	X_1^{t-1}		X_2^{t-2}	X_3^{t-3}
	<i>Age</i>		<i>Air pollution</i>	<i>Dirty water</i>
	b_0	$b_1, [\%]$	$b_2, [\%]$	b_3
1	91.767	97.505	84.814	4.4307
2	(0.88)	(25.3)	(5.0)	(7.5)
3	0.38	< 0.001	< 0.001	< 0.001
4	$R^2 = 91.8\%$ $R_{adj}^2 = 91.4\%$			$\sqrt{d} = 96.971$
5	$F = 228$ $v = 62$		< 0.001	
6	$d = 0.8202$	$\chi^2 = 5.95$	$D_n = 0.0745$	$\omega^2 = 0.0706$

Также во 2-й и 3-й строках здесь приводятся модули t -статистик при проверке $H_0: \beta_j = 0$ (для каждого j) и достигнутые уровни значимости (p -value); в строке 4 – коэффициент детерминации R^2 , его несмещенное значение (R_{adj}^2)² и несмещённая оценка стандарта $d^{0.5}$; в строке 5 – результаты дисперсионного анализа (статистика Фишера, число степеней свободы) и p -value для гипотезы $H_0: \beta_j = 0$, для всех $j = 1, \dots, 3$; наконец, в 6-й строке лежат значения четырёх статистик для критериев согласия распределения остатков, полученных в (1), с нормальным законом (Карлберг, 2018).

Средняя относительная ошибка аппроксимации (MRRE) была 2.7% при её расчёте по $n = 65$ регионам материала обучения. По определению, это есть среднее значение ошибок $\delta = 100 \cdot |Y_{est} - Y_f| / Y_f$, где Y_{est} – оценка Y по модели, Y_f – данные по факту из (ЗД-РФ, 2019). На этапе “экзамен”, MRRE получилась менее 5%; здесь по $n = 39$ регионам оценивалась заболеваемость Y^t в 2019 году, причём Y_f – данные из (ЦНИИОЗ, 2019). В список этих регионов, кроме почти всех перечисленных, вошли: Ненецкий АО; Владимирская, Новосибирская области; здесь экзамене не было лишь Архангельской и Тюменской областей (обе – без своих АО), Сахалинской области, Забайкальского края. Кроме того выполнялись проверки: “есть ли мультиколлинеарность?”, “имеется ли гомоскедастичность?”, “нет ли автокорреляции?”, “имеется ли нормальность остатков?”.

На основе (1) можно, используя лишь данные, которые ежегодно собираются Росстатом, оценить распространенность ЗНО среди населения многих регионов с горизонтом в 1-2 года. Это позволит более эффективно распределять ресурсы в сфере здравоохранения, в той или иной степени решать ряд задач национальных проектов по здравоохранению, экологии, демографии. Подтверждена важность поставки гражданам доброкачественной питьевой воды из сетей водоснабжения, необходимость борьбы в регионах страны с загрязнениями атмосферного воздуха, а также – борьбы за снижение сброса загрязнённых вод в поверхностные водные объекты.

Литература

- Айдинов Г.Т., Марченко Б.И. и др. Применение многомерных статистических методов при выполнении задач совершенствования информационно-аналитического обеспечения системы социально-гигиенического мониторинга // Здоровье человека и среда обитания. 2015. № 7 (268). С. 4-8.
- Бузинов Р.В., Кику П.Ф., Унгурану Т.Н. и др. От Поморья до Приморья: социально-гигиенические и экологические проблемы здоровья населения: монография. – Архангельск. 2016. 397 С. URL: <https://ecology.aonb.ru/ot-pomorya-do-primorya.html> (14.02.2021)
- ЗД-РФ (2019). Здоровоохранение в РФ: Стат.сб. / Приложение. – М.: РОССТАТ, 2019. 2015–2019.
- Карлберг К. (2018). Регрессионный анализ в Microsoft Excel / Пер. с англ. М., 2017. 400 с.
- Словарь (2003). Большой словарь географических названий, М., 2003. 832 с. ISBN 5-94799-148-9
- Сулейманов Р.А. и др. Анализ заболеваемости и смертности населения Республики Башкортостан злокачественными новообразованиями // Медицина труда и экология человека. 2019. № 2. С. 14–23. DOI: <http://doi.org/10.24411/2411-3794-2019-10016>
- Лазарев А.Ф. Петрова В.Д. и др. Многофакторный анализ при формировании групп высокого онкологического риска // Бюллетень медицинской науки. 2017. № 1 (5). С. 37–43. DOI: [http://doi.org/10.31684/2541-8475.2017.1\(5\).37-43](http://doi.org/10.31684/2541-8475.2017.1(5).37-43)
- Мирасова В.М., Малыгина Н.В. Определение зависимости уровня заболеваемости граждан регионов РФ от состояния окружающей среды с помощью многомерных статистических методов // "XXI век: итоги прошлого и проблемы настоящего плюс". 2017. 1(35).
- МНИОИ (2020). Злокачественные новообразования в России в 2019 г. (заболеваемость и смертность) – М.: МНИОИ им. П.А. Герцена, 2020. 252 с. ISBN 978-5-85502-260-5
- РР-СЭП (2020). Регионы России. Социально-экономические показатели. – М.: РОССТАТ. 2020.
- РР-ОХ (2019). Регионы России Основные характеристики субъектов Российской Федерации: Стат.сб. – М.: РОССТАТ. 2019. (<https://rosstat.gov.ru>).
- ООС-РФ (2018). Охрана окружающей среды в России: Стат.сб. М.: РОССТАТ. 2018. 2014–2018.
- ЦНИИОИЗ (2020). Социально-значимые заболевания населения России в 2019 году (статистические материалы). – М.: ЦНИИОИЗ МЗ РФ, 2020. 76 с. (www.mednet.ru)
- Фельдблюм И.В., Алыева М.Х., Радионова М.В. Комплексное влияние медико-социальных и средовых факторов риска на вероятность развития колоректального рака // ТМЖ. 2018. № 3 (73). С. 24-28. DOI: <https://doi.org/10.17238/PmJ1609-1175.2018.3.24-28>

ⁱ Степанов Владимир Сергеевич – ЦЭМИ РАН, Stepanov@cemi.rssi.ru