

С.Л. Плавинский

Моделирование ВИЧ-инфекции и других заразных заболеваний человека и оценка численности групп риска

Введение в математическую эпидемиологию

Число заболевших

Гиперэндемическая фаза
 $R > 1$

$R = 1$

Время

Изменение I
Эндемическая фаза
 $R < 1$

Москва
2009

С.Л. Плавинский

Моделирование ВИЧ-инфекции
и других заразных заболеваний человека
и оценка численности групп риска

Введение
в математическую эпидемиологию

Москва
2009



Данный документ разработан и издан по заказу Учебно-Консультационного Центра Открытого Института Здоровья в рамках проекта ГЛОБУС.

Плавинский С.Л.

Моделирование ВИЧ-инфекции и других заразных заболеваний человека и оценка численности групп риска. Введение в математическую эпидемиологию. — М., 2009. — 100 с.

Содержание

Введение	4
1. Математическое моделирование в эпидемиологии	12
2. Методы обратного расчета	15
2.1. Анализ заболеваемости по количеству манифестных форм	15
2.2. Обратный расчет с аналитической формой описания прогноза развития манифестных форм	19
3. Динамическое моделирование	28
3.1. Детерминистские модели	28
3.1.1. Создание графов динамики заболевания	28
3.1.2. SIS-модели	31
3.1.3. SIR-модели	49
3.1.4. Модели открытых популяций	54
3.1.5. Как улучшить модели?	61
3.2. Стохастические модели	62
4. Методы двойного охвата	75
4.1. Метод Линкольна-Петерсена	75
4.2. Методы с учетом корреляции источников	78
4.3. Методы для открытой популяции	86
Заключение	95
Список литературы	96
Приложение	98

Введение

Конец XX века в медицине вообще и медицинской науке в частности охарактеризовался тремя важными изменениями. Во-первых, в число важнейших заболеваний вернулись инфекционные болезни, причем не только в развивающихся, но и в развитых странах. Основной причиной этого возврата стала глобальная эпидемия ВИЧ/СПИД. Во-вторых, появилась доказательная медицина, направленная на использование в клинической практике только проверенных в эксперименте методов лечения и профилактики и, в-третьих, экономика здоровья оформилась как самостоятельная дисциплина, которая анализировала не просто эффективность, а стоимость-эффективность вмешательств, базируясь на положении о том, что ввиду ограниченности ресурсов здравоохранения неэффективное расходование средств означает невозможность оказать помощь больным людям.

Доказательная медицина и экономика здоровья стали дополнять друг друга, отвечая на вопрос, какие новые методы лечения и профилактики являются эффективными и, если они являются эффективными, то являются ли они стоимостно-эффективными.

Основным инструментом оценки эффективности и дальнейшего анализа стоимости-эффективности являются рандомизированные контролируемые испытания, эксперименты, в которых группа пациентов делится на две подгруппы, одна из которых получает новое, экспериментальное вмешательство, а другая является группой сравнения. Этот дизайн исследований считается наиболее адекватным для оценки эффективности методов лечения, поскольку позволяет регистрировать важные для пациентов исходы и ограничивает возможность систематических ошибок, связанных с неучтенными факторами риска среди пациентов.

Вместе с тем, в области инфекционных заболеваний у подобного подхода к оценке эффективности лечебных и в первую очередь профилактических мероприятий есть целый ряд недостатков. Одной из причин является заразный характер этих заболеваний. В случае неинфекционных заболеваний весь положительный эффект от вмешательства ограничивается тем, что наблюдается в исследовании. Однако в случае заразных заболеваний каждый предотвращенный случай заболевания может привести к предотвращению других случаев, формируя цепь положительных результатов, не видимых в исследовании из-за ограниченности популяции. С другой стороны, если заболевание распространено достаточно широко или связано с поведением риска, то предотвращение случая заболевания в рамках исследования не означает, что человек не инфицируется немедленно после его окончания. Таким образом, в случае инфекционных заболеваний стандартный подход с использованием рандомизированных контролируемых испытаний в качестве золотого стандарта для определения эффективности дает сбой.

Возможной альтернативой являются популяционные рандомизированные контролируемые испытания (т.н. кластерные рандомизированные испытания), в которых единицей рандомизации и наблюдения является не конкретный человек, а население определенного населенного пункта или региона. К сожалению, эти исследования достаточно дорогостоящие и также не свободны от ряда проблем, в частности, «перелива», когда профилактическое вмешательство в одном регионе оказывает воздействие на соседний (для чистоты эксперимента регионы/населенные пункты должны быть полностью отделены друг от друга, что в современном мире практически невозможно).

Наличие проблем с оценкой эффекта профилактических программ в области эпидемиологии инфекционных заболеваний человека привело к тому, что исследователи стали обращать внимание на возможность построения математических моделей распространения заразных заболеваний

и затем на этих моделях изучать последствия тех или иных вмешательств и сравнивать их с реально наблюдаемыми результатами.

Хотя построение моделей не может считаться полностью адекватной заменой эксперименту, тем не менее они позволяют достаточно адекватно предсказать ситуацию, если справедливы все использованные в модели допущения. И здесь кроется важнейшее правило, которое всегда должно использоваться при анализе любых моделей — любая модель лишь настолько хороша, насколько хороши заложенные в нее допущения.

Стоит отметить, что в реальности и упоминавшиеся выше рандомизированные контролируемые испытания (РКИ) являются моделями. И у них есть допущения (самое важное из них — генерализуемость или обобщаемость результата). Оценка результатов РКИ также включает использование математических моделей в виде статистических моделей, которые также базируются на целом ряде допущений. Однако в данном пособии мы будем говорить о математических моделях, отличных от статистических — моделях, которые были специально разработаны для изучения процесса распространения заразных заболеваний человека и в последнее время активно используются для изучения процесса распространения ВИЧ-инфекции и оценки эффективности профилактических мероприятий.

Мы не будем упоминать модели, которые изучают процессы эволюции заразного заболевания на индивидуальном уровне (например, математические модели эволюции вирусной популяции в организме или иммунного ответа [17]). Также не будут рассматриваться вопросы моделирования взаимодействия между популяцией человека и возбудителя [9]. Задача этого пособия значительно скромнее: познакомить читателя с некоторыми простейшими инструментами математического моделирования заразных заболеваний человека, которые можно использовать в повседневной практике.

Перед автором стояла задача дать не только теоретическое описание используемых в описываемой области методов, но и привести примеры, которые можно было бы самостоятельно опробовать, поскольку самое лучшее обучение — это обучение через действие. Однако в области современного моделирования это означает необходимость полагаться на определенные компьютерные пакеты, поскольку выполнение моделирования «вручную» хоть и возможно, но может отбить охоту к этому занятию даже у самого терпеливого исследователя. В последние годы были разработаны достаточно интересные и мощные пакеты математического моделирования, приспособленные для анализа ситуации с ВИЧ/СПИД, в частности, система Spectrum¹. Эта система достаточно широко использовалась в публикациях, посвященных динамике распространения ВИЧ-инфекции, однако она разрабатывалась как инструмент для выполнения специфических заданий и уже содержит ряд допущений, настолько тесно вплетенных в нее, что любой пользователь должен их принять, если хочет пользоваться этой системой. С другой стороны, эта система и ей подобные не очень афишируют используемые в моделях допущения, что способствует ложному чувству их отсутствия и простоты моделирования. Для целей этого пособия такой подход был неприемлем, поскольку целью являлось как раз продемонстрировать, как можно двигаться от простых моделей к более сложным, что при этом выигрывается, а что теряется. Поэтому было принято решение ограничиться использованием только двух инструментов для моделирования. Один из них — среда для визуального моделирования систем Vensim, которая является более чем адекватным инструментом для того, кто хочет познакомиться с моделированием систем. Однако и у нее есть ограничения, и поэтому многие примеры в пособии приведены в системе SAS, которая больше известна как система статистической обработки данных и как таковая широко распространена в медицинском мире. У SAS есть достаточно мощный язык и набор процедур, который позволяет описывать модели и анализировать их. При этом у аналитика остается полная свобода

¹ <http://www.policyproject.com/software.cfm?page=Software&ID=Spectrum>

модификации этих моделей и изменения допущений, на которых они строятся. Естественно, для детального овладения этими моделями необходимо знать язык SAS, однако он не является очень сложным, и изучение языка для проведения моделирования является неизбежным злом, поскольку визуальные методы построения моделей становятся мало приемлемыми в случае моделей, по сложности хоть немного напоминающих реальность.

В рамках этого предисловия хотелось бы вкратце познакомить читателя с идеями математического моделирования заразных заболеваний, в первую очередь ВИЧ/СПИД, не касаясь подробностей, которые будут изложены в дальнейшем.

Интерес к математическим моделям в инфекционной эпидемиологии возродился во второй половине XX века, когда эпидемиологи столкнулись с двумя проблемами, имевшими прямое и непосредственное отношение к эпидемии ВИЧ/СПИД: (1) как оценить количество случаев заболевших (ВИЧ-инфицированных), если известно только количество людей в терминальной (клинически манифестной) стадии заболевания (СПИД) и (2) как предсказать распространение инфекции в популяции на основании информации о ее заразности и частоте поведения риска. Ответ на первый вопрос привел к появлению методов обратного расчета, которые базировались на достаточно простой, но интересной идее — если нам известно количество лиц в терминальной стадии и мы знаем (или можем сделать предположение) о частоте перехода заболевания в терминальную стадию, то можно достаточно просто оценить, какое количество лиц должно быть в популяции для того, чтобы наблюдалось заданное количество лиц в клинически манифестной стадии. Строго говоря, для этого надо вначале выполнить расчеты в прямом направлении — если в год X у нас было Z инфицированных, то какое количество их них перейдет в клинически манифестную форму заболевания к году Y . Эту процедуру можно повторить для разных предполагаемых значений Z . Соответственно, зная количество лиц с манифестной формой заболевания в год Y , можно найти неизвестное количество инфицированных Z , заглянув в построенные таблицы. Именно такой обратный поиск и дал название методу.

Очевидно, что основной проблемой в данном случае является форма зависимости между количеством инфицированных и количеством манифестных форм. Эту информацию нельзя получить иным способом, кроме как наблюдением за инфицированными пациентами. Проведенные в начале эпидемии ВИЧ-инфекции исследования позволили получить подобные данные и предложить уравнения для описания зависимости, которые, как оказалось, относятся к достаточно хорошо известному среди статистиков распределению Вейбулла. Знание формы распределения позволило упростить расчеты, однако породило проблему переносимости данных (с тех пор именно распределение Вейбулла с параметрами, определенными в ходе тех исследований, используется в большинстве модельных программ). Действительно, насколько данные, полученные в США в начале 1980-х, применимы в России во втором десятилетии XXI века? На этот вопрос можно ответить, только анализируя предсказания, полученные при использовании модели, и сравнивая их с реальными данными.

Интерес к методам обратного расчета стал снижаться после появления высокоактивной антиретровирусной терапии, поскольку она сильно изменила течение ВИЧ-инфекции, и теперь невозможно уже стало использовать этот подход для оценки численности ВИЧ-инфицированных, поскольку форма изменилась. Однако разработанные подходы можно использовать на популяциях, в которых распространенность ВААРТ относительно невысока и можно использовать в качестве маркеров нынешнего состояния не только наличие СПИД, но и другие показатели, например, уровень CD4+ Т-лимфоцитов. Именно в этой области — оценки динамики эпидемического процесса в популяции на основании иммунологических маркеров — и развиваются сейчас методы обратного расчета.

Методы обратного расчета позволили ввести в обиход эпидемиологов инструменты математического моделирования и подготовили почву для принятия этих инструментов в попытках ответа на второй важнейший вопрос — как будет развиваться эпидемиологический процесс в будущем. На первом этапе ответ на этот вопрос являлся естественным следствием использования методов обратного расчета — описав динамику развития эпидемического процесса, можно было достаточно легко оценить, что будет происходить с частотой появления манифестных форм. Однако эпидемиологов интересовало не только описание процесса на уровне всей популяции, но и ситуация в отдельных группах риска, а также ответы на вопросы типа «а что будет, если...?». Эти ответы не могли быть даны на основе моделей обратного расчета, и поэтому появились новые (точнее, хорошо забытые старые) модели, относившиеся к большому семейству динамических моделей.

Динамические модели являются концептуально достаточно простыми. Популяция представляется как набор групп гомогенных объектов. Сложность модели зависит от количества этих групп. Простейшие модели состоят из одной группы. Внутри группы контакты между объектами считаются случайными (т.е. не зависящими от заразного статуса человека) и равновероятными (это допущение более справедливо для воздушно-капельных инфекций, нежели для инфекций с контактным механизмом передачи). Иными словами, как писал в романе «Мы» Е. Замятин: «Всякий из нумеров имеет право — как на сексуальный продукт — на любой номер». Мы пока оставим в стороне некоторую искусственность этого допущения, отметив, что можно начинать делить популяцию на все меньшие и меньшие группы до тех пор, пока оно не станет выполняться (именно это и делается в системах типа Episims [2]). Однако, если теперь взять такую гомогенную группу, то вероятность заражения в ней зависит от количества контактов и вероятности заражения при однократном контакте. Располагая этой информацией, можно описать процесс распространения заразного агента в этой группе. При этом возможны несколько сценариев развития событий — заболевание может приводить к появлению стойкого иммунитета, тогда количество контактов, при которых может произойти заражение, будет прогрессивно уменьшаться (по мере роста иммунной прослойки). Такое заболевание распространится по популяции (группе), а затем исчезнет. Эти модели называются моделями SIR (от английского ‘susceptible-infected-resistant’ — уязвимый-инфицированный-резистентный). Возможно, что переболев, человек снова может заразиться — тогда заболевание останется в популяции навсегда. При этом течение заболевания может быть волнообразным или со стабильным уровнем. Такие модели носят название SIS (от английского ‘susceptible-infected-susceptible’ — уязвимый-инфицированный-уязвимый). Посредине между этими двумя крайними типами моделей находятся модели SID (от английского ‘susceptible-infected-dead’ — уязвимый-инфицированный-мертвый), в которых предполагается, что после стадии инфицирования человек по какой-то причине удаляется из группы (как следует из названия этой модели, причиной удаления может быть смерть). В этом случае численность группы прогрессивно снижается, и ее судьба зависит от заразности заболевания и продолжительности жизни инфицированного человека (чем эти два параметра выше, тем выше вероятность исчезновения всей группы). На самом деле модели SID не очень адекватны в долгосрочной перспективе на основании эволюционных аргументов. Если заболевание приводит к «вымиранию» всей группы, то вместе с ней вымирает и возбудитель и, соответственно, такое заразное заболевание рано или поздно должно исчезнуть. Аналогичным образом с эволюционной точки зрения не вполне адекватны и SIR-модели. Однако заболевания, отвечающие допущениям SIR-моделей, однозначно существуют, почему же это так?

Ответ достаточно прост. Потому, что рассматривая выше все модели, мы предположили, что группа является замкнутой. Иными словами, не наблюдается ее динамики — в ней не появляются новые члены, а старые не уходят. Это называется допущением закрытой популяции. В реальности, естественно, ни одна популяция не является закрытой. В любой из них появляются новые члены, а старые исчезают. В таких условиях даже для моделей SIR или SID не наблюдается истощения количества уязвимых (за счет притока новых членов). Как можно показать (и это будет показано позднее), SIR- и SID-модели в условиях открытой популяции, т.е. популяции,

в которой могут появляться новые члены и из которой могут уходить старые, начинают вести себя аналогично моделям SIS. Именно поэтому SIS-модели играют важнейшую роль в изучении динамики инфекционных заболеваний.

Для SIS-моделей существует один важнейший показатель, который характеризует развитие заразного процесса. Этот показатель называется репродуктивным числом и определяется как среднее количество лиц, которых заразит один больной за время своей болезни. Различают два типа репродуктивных чисел — базовое репродуктивное число, которое оценивает среднее количество зараженных при условии попадания больного в полностью восприимчивую популяцию, и просто репродуктивное число, которое отражает течение эпидемического процесса в данный момент времени. Легко показать, что репродуктивное число равно произведению базового репродуктивного числа на процент уязвимых в популяции. Базовое репродуктивное число описывает судьбу эпидемического процесса в случае заноса заболевания в популяцию, тогда как репродуктивное число указывает на тенденции в данный момент времени. Если это число оказывается выше единицы, количество инфицированных будет нарастать. Если оно ниже единицы — то количество инфицированных снижается. Если репродуктивное число равно единице, количество заражающихся и выздоравливающих уравниваются и инфекция персистирует в этой группе.

Базовое репродуктивное число оказывается равным произведению вероятности заражения при однократном контакте на количество контактов за время болезни, что, в свою очередь, равно произведению количества контактов в единицу времени на продолжительность заболевания. Таким образом, эпидемический процесс оказывается зависимым от особенностей возбудителя (вероятность заражения и длительность заразного периода) и поведения хозяина (количество контактов). Знание несложного уравнения, описывающего репродуктивное число, позволяет планировать профилактические мероприятия по всем направлениям:

- как они влияют на вероятность заражения (для профилактики ВИЧ примером является использование презервативов);
- как они влияют на частоту контактов (для профилактики ВИЧ-инфекции это означает избегать случайных половых связей, А и В в подходе АВС²);
- как они влияют на длительность заболевания — вернее, длительность заразного периода (для профилактики ВИЧ-инфекции это может быть дополнительным обоснованием необходимости антиретровирусной терапии).

Это уравнение также подчеркивает необходимость всеобъемлющего подхода к профилактическим и противоэпидемическим мероприятиям, использования всего комплекса мероприятий, и позволяет оценивать необходимые целевые показатели, если базовое репродуктивное число и его составляющие известны. С другой стороны, важность базового репродуктивного числа для определения динамики эпидемического процесса подчеркивает важность знания для данной группы лиц его компонентов (частоту контактов, вероятность заражения и т.д.), поскольку они позволяют оценить потенциал для развития заболевания в этой группе. Именно поэтому данные параметры поведения зачастую являются основой дозорного эпиднадзора в группах риска. Использование даже простейших моделей на основе базового репродуктивного числа зачастую показывает ограниченность знаний эпидемиологов и специалистов общественного здоровья о возможностях и путях распространения инфекции, стимулируя к проведению исследований, которые бы устранили имеющийся дефицит знаний. Если бы единственным результатом

² АВС — трехступенчатая модель профилактики инфекций, передающихся половым путем, состоящая из предложения использовать А (abstinence) — воздержание до брака, В (be faithful) — моногамные половые отношения в браке или партнерстве и С (condom) — использовать презервативы в случае смены партнера или наличия случайных партнеров. В русском языке часто используется сокращение ВВП (воздержание, верность, презерватив).

использования моделей было указание на то, какие данные следует искать для прогнозирования эпидемической ситуации, то они уже не были бы зря разработаны.

Знание базового репродуктивного числа позволяет оценить и уровень персистирования инфекции в популяции. Легко показать, что в случае SIS-модели количество уязвимых в стационарной фазе эпидемического процесса (достижения эндемического уровня заболеваемости) будет равно величине, обратной базовому репродуктивному числу (т.е. если базовое репродуктивное число 1.1, то количество уязвимых составит 91%, а, соответственно, количество инфицированных — 9%). Это простое правило позволяет оценить, дошел ли процесс в группах риска до уровня «насыщения», а с другой стороны, позволяет оценивать базовое репродуктивное число в группах, где, как видно по результатам эпидемиологического исследования, распространенность вышла на плато. Это, в свою очередь, позволяет получать информацию о тех компонентах уравнения базового репродуктивного числа, которые напрямую не могут быть измерены (или измерение осложняется, как, например, измерение количества лиц, с которыми потребитель инъекционных наркотиков совместно использует шприцы и иглы). Базовое репродуктивное число может также использоваться для оценки потребности в покрытии вакцинацией, если вакцина от данного заболевания существует.

В том случае, если популяция состоит из гетерогенных групп, были разработаны методики, позволяющие оценить суммарное репродуктивное число либо на основе дисперсии (разброса) показателя в популяции (например, численности контактов), либо путем оценки репродуктивных чисел в подгруппах и расчета средневзвешенного репродуктивного числа в зависимости от количества контактов, приходящихся на эту группу. Полученные таким образом оценки репродуктивного числа могут использоваться, как описано выше, для установления целевых показателей для противоэпидемических мероприятий и предсказания эволюции эпидемического процесса в будущем.

В целом анализ динамических моделей, который дополняется анализом репродуктивного числа, является достаточно полезным инструментом для специалистов общественного здоровья, занимающихся планированием профилактических и противоэпидемических мероприятий в области противодействия распространению заразных заболеваний человека.

Динамические модели, особенно если они включают разумно большое количество относительно однородных групп и описывают открытую популяцию, позволяют достаточно хорошо моделировать распространение заразных заболеваний и предсказывать их динамику. Они могут служить основой для оценки эффективности мероприятий в рамках подходов, принятых в экономике здоровья, поскольку хорошо откалиброванная модель позволяет выйти за рамки профилактического исследования и проанализировать различные сценарии развития событий, оценить их стоимость и эффект для здоровья населения. Таким образом может быть выполнена задача отбора наиболее стоимостно-эффективных вмешательств. У них есть только один серьезный недостаток — они справедливы для очень больших популяций.

Не является секретом, что очень часто точно предсказать динамику эпидемического процесса невозможно. В одной группе заболевание дает вспышку, в другой — аналогичной по всем параметрам — нет. Подобная ситуация не может быть описана в рамках обсуждавшихся выше моделей, поскольку они по своей природе являются детерминистскими — т.е. не оставляют место случайности. Однако в реальной жизни случайность играет очень большую роль, и поэтому целый раздел математической эпидемиологии посвящен стохастическим моделям, моделям, которые базируются на предположении о том, что в процессе распространения заразных заболеваний случай играет очень большую роль. Надо отметить, что, чем меньше популяция, для которой строится модель, тем выше важность стохастических моделей. Чем популяция больше, тем меньше будет вклад случайности в определение суммарных показателей и тем ближе будут

к истинным предсказаниям детерминистских динамических моделей. Поэтому можно сказать, что стохастические модели — это модели для маленьких групп и небольших популяций.

К сожалению, стохастические модели уже не могут предсказать, как именно будет развиваться эпидемический процесс в популяции. Они, собственно говоря, и построены на предположении, что это невозможно, что исключить влияние случайных факторов в распространении возбудителя нельзя. Однако эти модели позволяют оценить, какие сценарии развития событий являются более вероятными, а какие — нет. Строго говоря, этой информации уже достаточно для того, чтобы специалист общественного здоровья мог принимать решения в рамках известной методологии управления рисками. Чем выше вероятность развития данного сценария и больше его негативная значимость, тем больше усилий надо предпринять для того, чтобы избежать такого развития событий.

Кроме того, стохастические модели могут позволить оценить возможное распределение продолжительностей времени вспышки, количества зараженных во время вспышки и ряд других важных для принятия решения параметров.

В настоящий момент большинство программных пакетов для динамического моделирования не поддерживают стохастических моделей, и это также явилось одной из причин, почему в данном пособии в качестве базового программного пакета была выбрана система SAS, которая позволяет анализировать как детерминистские, так и стохастические модели.

Строго говоря, все другие используемые в математическом моделировании инфекционных болезней подходы могут быть сведены к описанным выше двум основным классам моделей — детерминистским и стохастическим, а внутри них они могут описываться моделями SIS, SID или SIR. Однако, в особенности в случае ВИЧ-инфекции, для повышения надежности моделирования часто не хватает очень важного параметра — размеров той популяции, для которой строится модель и в которой изучается динамика эпидемического процесса. А, как указывалось выше, это важно по трем причинам — во-первых, малые популяции требуют стохастического моделирования, во-вторых, популяционное воздействие распространения инфекции в группе зависит от относительных размеров этой группы (в конце концов, планирование расходов системы здравоохранения зависит от абсолютного числа лиц, нуждающихся в помощи) и, наконец, даже детерминистские динамические модели с несколькими группами требуют оценки относительной численности этих групп для описания процесса перехода инфекции из одной группы в другую.

По этим причинам в данное пособие были включены разделы, посвященные определению численности популяции, если полностью пересчитать объекты в этой популяции по каким-то причинам сложно. Эти методы, несмотря на то, что они тоже базируются на моделях, обычно не включаются в разделы по математическому моделированию и, в принципе, мало используются в эпидемиологии, несмотря на их широкое распространение в биологии вообще. Биологи давно сталкивались с проблемой подсчета количества особей в популяциях, которые не горят желанием контактировать с исследователем. Поэтому они разработали методы, которые по-английски называются *capture-recapture* (поймать, еще раз поймать), а по-русски их принято называть методами двойного охвата, хотя иногда охват является тройным или даже большим. Основная идея этого метода заключается в том, что, если мы можем каким-то образом пометить объекты из достаточно гомогенной популяции, затем вернуть их обратно в популяцию и снова найти несколько объектов, то количество ранее встреченных объектов во второй выборке будет — очевидно — тем меньше, чем больше размер популяции. Соответственно, можно разработать формулы, которые позволяют оценить численность всей популяции на основании численности «уже встречавшихся» объектов во второй выборке (численность популяции равна количеству лиц во второй выборке, деленному на процент лиц во второй выборке, которые уже встречались

в первой). Эти формулы, которые впервые были использованы для оценки численности популяции рыб и зверей, сейчас находят все большее применение для оценки численности лиц, занимающихся коммерческим сексом, потребляющих наркотики и относящихся к другим группам риска.

Использование этих простейших формул сопряжено с двумя типами проблем. Во-первых, предполагается, что первая встреча никак не влияет на вероятность второй встречи. Это зачастую не так: люди, столкнувшиеся первый раз с исследователем могут, захотеть прийти на вторую встречу (если за это дается какая-то награда), либо наоборот, избегать этой встречи (если она сопряжена с наказанием, например, арестом). Поэтому значительные усилия были приложены для выработки методологии учета возможности корреляции (связи) между попаданием в первую и вторую выборки. Все эти методы требуют большего количества выборок (или большего количества источников данных, которые позволяют найти представителей изучаемой группы), минимальным является количество в три выборки, для которых разработана относительно простая методология, базирующаяся на оценке этой корреляции с помощью стандартных статистических методов. Более сложные, но и более надежные методы базируются на построении статистических моделей, таких, как логлинейные модели, хорошо известные в биостатистике. Не удивительно, что эти расчеты очень удобно проводить в статистических системах, таких как SAS. Второй тип проблем связан с тем, что простые методы двойного охвата базируются на предположении закрытости популяции. Это предположение обосновано, если две выборки берутся достаточно близко друг к другу по времени. Однако в ряде случаев взять выборки близко по времени друг к другу не представляется возможным, а иногда исследователя интересует как раз динамика численности популяции риска.

В этом случае ему на помощь приходят модели для открытых популяций, первые из которых были разработаны в биологии для ориентировочных оценок еще в 50-х годах XX века. Эти методики требуют небольшого количества последовательных выборок (три), но они позволяют лишь прикинуть, остается ли популяция стационарной, или она растет или сокращается. Более сложные модели, разработанные в последние десятилетия, используют индивидуальную информацию о членах группы риска для того, чтобы оценить численность популяции и ее динамику на протяжении любого периода времени. Эти методы достаточно интенсивны с точки зрения расчетов и, хотя простейшие примеры могут быть проанализированы и вручную, сколько-нибудь реальные задачи требуют для своего решения помощи в виде компьютерных программ. И алгоритмы для подобного анализа разбираются в данном пособии.

Таким образом, современная эпидемиология может в значительной степени выиграть от использования методов математического моделирования в области определения истории эпидемического процесса, когда напрямую его наблюдать было сложно (методы обратного расчета), от определения размеров групп риска (методы двойного охвата) и использования полученных данных для моделирования эпидемического процесса либо на больших популяциях (детерминистское динамическое моделирование), либо в малых группах (стохастическое моделирование). В целом это использование математических моделей в эпидемиологии называют математической эпидемиологией. Введению в методы математической эпидемиологии и посвящено это пособие.

1. Математическое моделирование в эпидемиологии

Одной из основных задач современной эпидемиологии является не только констатация современного состояния проблемы инфекционной заболеваемости и выработка мероприятий по борьбе с существующими заразными болезнями человека, но и определение тенденций в развитии этих процессов.

Повторяющийся характер эпидемий в человеческой популяции уже давно позволял предполагать, что подобный анализ тенденций и предсказание наиболее вероятного пути развития ситуации возможны. Математическое моделирование инфекционных заболеваний имеет достаточно длительную историю. Обычно считается, что первой попыткой такого рода было создание Даниэлем Бернулли в 1760 году модели для изучения возможностей вакцинопрофилактики оспы. Однако реальный прогресс в этой области был достигнут лишь в XX веке, когда Хамер (1906), Росс (1911, 1916) и Кермак и МакКендрик (1927) сформулировали модели, базировавшиеся на принципе действующих масс и предположении о гомогенности скрещивания [13]. Особое влияние на дальнейшее развитие эпидемиологии оказала концепция репродуктивного числа, предложенная Россом в 1910 году при изучении распространения малярии.

Однако быстро стало понятно, что процесс развития эпидемического процесса вряд ли может быть чрезвычайно прост, поскольку он зависит от значительного числа различных факторов. Анализ многопараметрических моделей до появления персональных компьютеров был затруднен, что и ограничивало широкое распространение метода моделирования.

Кроме того, во второй половине XX века инфекционная заболеваемость стала снижаться, а одновременно стала снижаться значимость ее прогнозирования и необходимость в разработке моделей.

С появлением персональных компьютеров ситуация стала упрощаться. Теперь от аналитика требовалось только адекватно описать модель и снабдить ее необходимыми параметрами, остальное за него делала специализированная компьютерная программа. Надо заметить, что, конечно, специализированные компьютерные программы не создаются только для задач моделирования заразного процесса. Они являются примером программного продукта для динамического моделирования, с помощью которого можно изучать различные процессы — от демографии до организации бизнеса. Именно бизнес-пользователи и были первоначально движущей силой, которая способствовала разработке соответствующих программных пакетов.

Однако в начале 80-х годов мир столкнулся с новым инфекционным заболеванием — инфекцией, вызванной вирусом иммунодефицита человека (ВИЧ-инфекцией), на фоне подъема заболеваемости от старых противников (венерических заболеваний и туберкулеза). Системе здравоохранения необходимо было планировать выделение средств на борьбу с этими заболеваниями и, поскольку это надо было делать на будущее, динамическое моделирование заразного процесса было единственным методом, при помощи которого можно было оценить тенденции.

На самом деле оценивать тенденции в развитии заразного процесса можно путем двух подходов — экстраполяции и динамического моделирования. Экстраполяция опирается на данные по заболеваемости инфекционным заболеванием, собранные на протяжении относительно длительного периода времени. Предполагается, что, как процесс развивался в прошлом, так он будет развиваться и в будущем. Если в популяции наблюдаются волнообразные подъемы заболеваемости, то, определив период повторяемости подъемов, можно предсказать, когда наступит следующий. Самым популярным является метод разложения по Фурье, который позволяет описать, строго говоря, практически любой периодический процесс.

Однако для целей анализа инфекционной заболеваемости, с которой мир столкнулся в начале 80-х, метод экстраполяции в своей простейшей форме был неприемлем просто потому, что исторических данных о заболеваемости, например, ВИЧ-инфекцией, не было (это было новое заболевание), а для венерических заболеваний и туберкулеза они были крайне мало полезными, ибо до этого наблюдалось стабильное снижение показателей.

Было очевидно, что предсказание при помощи экстраполяции является ошибочным, поскольку рост заболеваемости связан с изменениями, происходящими в человеческом обществе, а метод экстраполяции предполагает, что все факторы, которые вызывают изменения заболеваемости, являются постоянными.

Одним из путей преодоления этой проблемы была сегрегация разных групп населения с разным поведением и анализ тенденций в отдельных подгруппах, а затем определение суммарных показателей. Например, если в популяции в целом венерических заболеваний было мало, а среди проституток их распространенность была постоянно высокой, то увеличение численности последних должно было приводить к росту общих показателей заболеваемости.

Еще одним возможным подходом является измерение ряда ключевых показателей и изучение их влияния на показатели заболеваемости с выработкой регрессионного уравнения, которое могло бы составить функцию, описывающую изменение процесса в зависимости от изменений не только времени, но и этих показателей. Подобный анализ является частью стандартного статистического процесса изучения распространенности заболеваний (обычно при помощи Пуассоновой регрессии, параметрических методов анализа выживаемости или модели Кокса).

Материал для подобного анализа достаточно сложно собрать и, на самом деле, методы анализа выживаемости и не очень предназначены для задач прогнозирования заболеваемости в последующий временной период. Проблема заключается в том, что значения показателя в каждый момент времени зависят от значений в предыдущий момент времени. Говорят, что такие данные коррелируют сами с собой (автокорреляция). Эта автокорреляция не позволяет пользоваться обычным методом анализа регрессионных моделей (метод наименьших квадратов) и требует иных подходов, которые иногда обозначаются как методы авторегрессии.

Однако данных не то что по распространенности болезней, но даже и численности подобных труднодоступных групп (проститутки³, гомосексуалисты⁴, наркоманы⁵), просто не было, и поэтому простота анализа полностью перечеркивалась отсутствием критических для анализа данных. Кроме того, как подчеркивалось выше, инфекционные заболевания часто носят волнообразный характер, что требовало использования нелинейных статистических моделей для аппроксимации. Нелинейные модели, во-первых, значительно сложнее в описании и подгонке, чем линейные, а во-вторых, требуют для анализа еще больше исторических данных. Причем желательно, чтобы оценки заболеваемости были получены на достаточно большой группе людей (точность определения показателей заболеваемости обратно пропорциональна квадратному корню из численности группы, в которой были зарегистрированы случаи заболевания).

³ В последнее время чаще используются имеющие меньший моральный оттенок термины «лица, оказывающие секс-услуги за плату» или «коммерческие секс-работницы» (КСР).

⁴ На самом деле в эту группу часто включают и лиц, вступающих в ситуационные гомосексуальные контакты, но не имеющих гомосексуальной ориентации, поэтому данную группу правильнее называть «мужчины, имеющие сексуальные контакты с мужчинами» (МСМ).

⁵ Для целей моделирования не важно, имеется ли у человека заболевание с непреодолимой тягой к использованию наркотических средств (наркомания) или он использует их по каким-либо другим причинам. Поэтому более адекватным термином является «потребитель наркотиков», а для анализа распространения гемоконтактных инфекций «потребитель инъекционных наркотиков» (ПИН).

Все эти факторы делали моделирование при помощи экстраполяции практически неприемлемым для использования, хотя, стоит заметить справедливости ради, они все равно использовались. В начале эпидемии анализировались тенденции распространенности синдрома приобретенного иммунодефицита (СПИД) в различных группах и затем делались прогнозы на несколько лет вперед. После того, как распространенность ВИЧ-инфекции в ряде групп достигала плато, метод переставал работать. Еще одним вариантом было изучение характерных эпидемических кривых в одних регионах и затем экстраполяция полученных данных на другие регионы (например, если выясняется, что в Сан-Франциско рост заболеваемости продолжался 8 лет, а затем достиг плато на таком-то уровне, то предполагается, что в Нью-Йорке, где он ниже, он будет продолжать расти до тех пор, пока не достигнет такого же уровня). К сожалению, этот метод базируется на неадекватном предположении о том, что в разных регионах заболевание распространяется в одинаковых по численности группах и с одинаковой скоростью.

2. Методы обратного расчета

Поскольку ВИЧ-инфекция была новым заболеванием, которое проявилось и стало регистрироваться по своему клиническому проявлению — СПИД, то для прогнозирования развития ситуации стали использовать методику обратного расчета. Эта методика базируется на том факте, что количество новых случаев СПИД в данный период времени зависит от количества случаев инфицирования ВИЧ в прошлом и распределения времени от момента заражения до развития СПИД. Если бы заболеваемость ВИЧ-инфекцией и распределение времени от момента заражения до наступления СПИД было известно, то определение заболеваемости СПИД просто сводилось бы к суммированию:

$$A(t) = \int_0^t g(s) F(t-s) ds,$$

где $A(t)$ — кумулятивное количество случаев СПИД до времени t , $g(s)$ — это скорость инфицирования в календарное время s (то есть количество новых случаев инфицирования в единицу времени s) и $F(t|s)$ — это распределение периодов инкубации для индивидуума, инфицированного в календарное время s . Распределение инкубационных периодов — это вероятность того, что инкубационный период будет короче времени t .

Обратный расчет — это попытка оценить заболеваемость ВИЧ-инфекцией и время от момента заражения до развития СПИД на основании исторических данных по заболеваемости СПИД. Обычно расчеты проводятся по отдельным группам риска. Использование этого метода позволило продемонстрировать, что использование антиретровирусной терапии приводит к удлинению периода от заражения до развития СПИД (были проведены расчеты, предполагавшие, что распределение времени от заражения до развития СПИД не меняется, которые показали, что реальное количество случаев СПИД меньше расчетного. Использование допущения нестационарного времени от заражения до развития СПИД (удлинение под влиянием терапии) позволило устранить этот кажущийся парадокс).

2.1. Анализ заболеваемости по количеству манифестных форм

Разобраться с техникой метода обратного расчета в принципе не сложно. Сделаем это на простейшем примере. Предположим, что в определенный год заразилось 100 000 человек. Однако, поскольку само заражение не приводит к клиническим проявлениям, ни пациент, ни система здравоохранения эти 100 000 человек зарегистрировать не могут. Однако по прошествии определенного времени у зараженного человека появляется клинически выявляемая форма болезни (очевидно, что речь может идти, например, о ВИЧ-инфекции и СПИД, соответственно). Наблюдение за группой недавно инфицировавшихся людей показывает, что каждый год постоянный процент лиц переходит в клинически видимую форму⁶. Подобное распределение времени от момента заражения до наступления клинически выявляемой формы заболевания описывается экспоненциальным уравнением:

$$y(t) = 1 - e^{-\frac{t}{k}},$$

⁶ Установление формы распределения времени от момента заражения до наступления клинически явной формы заболевания, а также оценка параметров этого распределения являются одними из наиболее сложных разделов в оценке заболеваемости методами обратного расчета.

где $y(t)$ — процент лиц, у которых развилось клинически выраженное заболевание ко времени t , а k — постоянная величина, прямо пропорциональная медианной продолжительности инкубационного периода⁷, и e — основание натуральных логарифмов.

Если продолжительность инкубационного периода t составляет, например, 8 лет⁸, то система здравоохранения будет наблюдать следующие количества лиц с вновь выявляемой клинически явной формой в года, следующие за заражением (см. табл. 1).

Таблица 1

Количество лиц, у которых разовьется заболевание в года, следующие за заражением

Года после заражения	Процент заболевших (%)	Абсолютное количество
1	8,30	8 300
2	15,91	15 910
3	22,89	22 889
4	29,29	29 289
5	35,16	35 158
6	40,54	40 540
7	45,47	45 475
8	50,00	50 000
9	54,15	54 150
10	57,96	57 955

Обратите внимание, что система здравоохранения наблюдает только значения в последней колонке табл. 1, исходное число заразившихся ей не известно. Однако, зная закон распределения времени от момента заражения до развития клинически манифестной формы, можно утверждать, что количество лиц с клинически манифестной формой заболевания в год t является суммой произведения вероятности развития манифестной формы заболевания за t лет на количество лиц, заразившихся t лет назад, произведения вероятности развития манифестной формы заболевания за $t-1$ год на количество лиц, заразившихся $t-1$ год назад и так далее. Например, в случае экспоненциального распределения времени заражения с медианным инкубационным периодом 8 лет, количество лиц с манифестной формой заболевания на 3-й год развития эпидемического процесса будет равно 0,2289, умноженное на количество лиц, заразившихся 3 года назад, плюс 0,1591, умноженное на количество лиц, заразившихся два года назад, плюс 0,083, умноженное на количество лиц, заразившихся год назад. Для описания наблюдаемых данных, например, за 6 лет, можно составить таблицу, аналогичную табл. 2.

Таблица 2

Компоненты наблюдаемой заболеваемости за первые 6 лет наблюдения

Год	Число заразившихся						Наблюдаемое число
	n_1	n_2	n_3	n_4	n_5	n_6	
1	0,083	0	0	0	0	0	8 300
2	0,159	0,083	0	0	0	0	15 910
3	0,229	0,159	0,083	0	0	0	22 889
4	0,293	0,229	0,159	0,083	0	0	29 289
5	0,352	0,293	0,229	0,159	0,083	0	35 158
6	0,405	0,352	0,293	0,229	0,159	0,083	40 540

⁷ Величина k равна медианной продолжительности инкубационного периода (t), деленной на $-\ln(0,5)$.

Даже беглый взгляд на табл. 2 позволяет понять, что в ней описана простейшая задача решения системы линейных уравнений. В данном случае — система из шести уравнений с шестью неизвестными ($n_1 - n_6$). Так, первые два уравнения этой системы будут выглядеть так:

$$8\ 300 = 0,83 * n_1$$

$$15\ 910 = 0,159 * n_1 + 0,83 * n_2$$

Решение подобной системы уравнений достаточно просто, и мы проиллюстрируем, как это можно сделать в системе статистического анализа SAS.

В SAS система линейных уравнений может быть решена целым рядом процедур, однако в данном случае воспользуемся процедурой REG.

Прежде всего надо ввести данные в систему. Сделать это можно следующим образом⁹:

1. DATA coeff;
2. ARRAY nn {10} nn1–nn10;
3. k=–8/LOG(0.5);
4. DO year=1 TO 10;
5. DO j=1 to 10;
6. if (year–j+1)>0 THEN nn [j]=1–exp(–(year–j+1)/k);
7. ELSE nn[j]=0;
8. END; OUTPUT;
9. END;
10. RUN;

Этот программный код создает массив из 10 элементов (по числу лет с неизвестным количеством вновь заразившихся). Затем два вложенных цикла заполняют массив либо значениями процента лиц, у которых к этому году разовьется манифестная форма заболевания, либо нулем, если соответствующая группа еще не заразилась (так, например, для тех, кто заразился на третьем году эпидемии, вклад в количество манифестных случаев первого и второго годов эпидемии является нулевым). Коэффициент k рассчитывается в программе так, чтобы его было легко менять в зависимости от наблюдаемой медианной продолжительности инкубационного периода¹⁰.

Далее в систему следует ввести наблюдаемое количество случаев манифестных форм заболевания, например, следующим образом:

1. DATA cases;
2. INPUT year cases;
3. CARDS;
4. 1 8300

⁸ При анализе временных показателей средними не пользуются, поскольку достаточно одного человека, у которого не развилось заболевание и инкубационный период будет увеличиваться каждый год, тогда как медиана — продолжительность времени, за который развилось клиническое заболевание у 50% обследованных — не меняется.

⁹ Нумерация строк в приводимых ниже программах SAS дается исключительно для удобства читателей и в реальном коде не используется. Кроме того, подчеркиванием и жирным шрифтом выделены основные команды SAS.

¹⁰ Обратите внимание, что в системе SAS для расчета натуральных логарифмов \ln используется функция LOG().

5. 2 15910
6. 3 22889
7. 4 29289
8. 5 35158
9. 6 40540
10. 7 45475
11. 8 50000
12. 9 54150
13. 10 57955
14. ;
15. RUN;

Теперь можно объединить два созданных файла в один для анализа процедурой регрессии:

1. PROC SORT DATA=coeff; BY year; RUN;
2. PROC SORT DATA=cases; BY year; RUN;
3. DATA dat_back;
4. MERGE cases coeff;
5. BY year;
6. RUN;

Оба файла были для объединения вначале отсортированы по переменной year (год анализа), а затем объединены командой MERGE. В результате получается файл с информацией, аналогичной представленной в табл. 2.

Анализ этого материала проводится очень просто:

1. PROC REG DATA=dat_back;
2. MODEL cases=nn1–nn10/NOINT;
3. RUN;

Приведенный код вызывает процедуру регрессионного анализа (PROC REG), моделирующую наблюдаемое количество манифестных случаев как функцию неизвестного числа заразившихся за время наблюдения (nn1– nn10). В качестве коэффициентов для определения того, какая пропорция заразившихся будет иметь в данный год манифестную форму, выступают рассчитанные ранее экспоненциальные коэффициенты. Опция NOINT подавляет внесение в модель постоянного члена (приравнивает его к нулю), поскольку в описанных выше уравнениях он не предусмотрен¹¹. В результате система выдает следующее решение:

¹¹ Логически постоянный член уравнения интерпретируется как число лиц с манифестной формой заболевания в отсутствие зараженных — число, очевидно равное нулю.

Переменная	Оценки параметра				
	Число степеней свободы	Оценка параметра	Стандартная ошибка	t-значение	Pr > t
nn1	1	100 005	.	.	.
nn2	1	-13,65800	.	.	.
nn3	1	7,22000	.	.	.
nn4	1	2,75655	.	.	.
nn5	1	2,09798	.	.	.
nn6	1	1,24428	.	.	.
nn7	1	-3,80453	.	.	.
nn8	1	-4,99968	.	.	.
nn9	1	6,70761	.	.	.
nn10	1	-6,82900	.	.	.

Анализ показывает, что большинство заразилось в первый год (что и было условием в данном примере). Однако незначительные величины есть и у случаев заражения в последующие годы (иногда коэффициенты являются отрицательными). В нашем случае это является следствием округления, однако в реальных условиях подобные колебания предсказанного числа инфицированных могут определяться также и случайными факторами (колебания выявляемости, другие небольшие колебания). Вместе с тем очевидно, что в данном случае колебания составляют доли процента (максимальное отклонение от истинно нулевого значения составляет 0,013%). Подобная благоприятная ситуация может и не сложиться, если количество инфицированных относительно небольшое и случайные колебания оказывают большее воздействие. Отсюда следует вывод, что методики обратного расчета очень чувствительны к количеству наблюдений и, чем их больше, тем точнее полученные с их помощью результаты.

Получив информацию о количестве зараженных в предшествующие годы, можно экстраполировать полученную информацию на будущее, сделав предположение о том, что в последующие годы заражения не происходит. Этот «оптимистичный» сценарий позволяет определить минимальные потребности системы здравоохранения, например, в лекарствах для терапии ВИЧ/СПИД. Так, в приведенном выше примере было установлено, что все случае манифестной формы заболевания являются следствием однократного заражения 100 000 человек десять лет назад. Зная это и используя описанное ранее уравнение, на 11-м году будет 61 445 лиц с манифестной формой заболевания, или на $(61\ 445 - 57\ 955) = 3\ 490$ человек больше, чем на 10-м году.

2.2. Обратный расчет с аналитической формой описания прогноза развития манифестных форм

Вместе с тем даже из приведенного выше примера понятно, что предположение о независимости количества инфицированных в год t от числа заразных лиц в год $t-1$ приводит к резкой потере точности. Представим себе, что можно описать процесс развития эпидемии¹² простой формулой, например — кривой экспоненциального роста (что возможно, например, на ранних стадиях развития процесса):

$$N(t) = N_0 * e^{t*m},$$

где $N(t)$ — количество лиц, инфицированных в год t (ранее, когда количество инфицированных в данном году не зависело от количества в прошлом, мы обозначали это значение n_t , см. табл. 2),

¹² В данном случае речь идет о заражении, а не о развитии манифестных форм.

N_0 — исходное количество зараженных, t — количество лет, прошедших с момента начала эпидемии, а m — постоянный коэффициент, показатель скорости роста эпидемии.

Тогда количество лиц, инфицированных в год t , будет зависеть от двух параметров, вне зависимости от того, за сколько лет у нас есть статистические данные. Соответственно, имея, например, данные за 10 лет, речь будет идти не о 10 уравнениях с десятью неизвестными, а о 10 уравнениях с 3 неизвестными (исходного количества инфицированных N_0 , скорости роста эпидемии m и продолжительности инкубационного периода t). В этом случае, анализ может принять во внимание случайные колебания, и результат окажется более адекватным.

В реальности обычно эпидемический процесс растет по экспоненциальному закону только короткий промежуток времени, поэтому обычно кривую заболеваемости описывают либо квадратичным уравнением (см., например, [22]), либо экспоненциальным уравнением с квадратичными членами [19]:

$$N(t) = N_0 * e^{a_1*t - a_2*t^2},$$

где a_1 и a_2 — неизвестные нам параметры роста эпидемии, которые надо оценить.

Описываемый квадратичным (или квадратично-экспоненциальным) уравнением процесс имеет фазу быстрого роста, а затем замедления (если коэффициент a_2 положителен). В долгосрочной перспективе заболеваемость, описываемая этим уравнением, сходит на ноль, что соответствует динамике многих заразных заболеваний. Однако требование снижения заболеваемости не является обязательным. Если коэффициент a_2 оказывается близким к нулю, то заболеваемость может расти на протяжении значительного периода времени. Таким образом, подобное уравнение позволяет описать большую часть возможных траекторий развития эпидемии.

Для того, чтобы выяснить, чему равны коэффициенты a_1 и a_2 , необходимо провести моделирование. Данные для него можно получить на основании наблюдения за когортой лиц, первоначально не инфицированных (и отмечать новые случаи заболеваний), либо за счет анализа данных по регистрации ВИЧ [19]. В последнем случае предполагается, что зарегистрированы не все данные, и поэтому полученное при моделировании значение исходной численности больных (N_0) в дальнейшем не используется, а используется только информация о форме кривой траектории эпидемии (предполагая, что данные по регистрации правильно оценивают тенденции в заболеваемости, хотя и могут занижать число инфицированных). Как выполняется подобный анализ, мы рассмотрим позднее на конкретном примере, а пока можно предположить, что в результате анализа было установлено, что $a_1=2,5$, а $a_2=0,2$. Если бы в стране со 100 инфицированными появилось заболевание с такими параметрами динамики заболеваемости и с медианным периодом до развития манифестной формы 8 лет, то наблюдаемое количество случаев (манифестной формы) выглядело бы так, как показано в табл. 3.

Для анализа этих данных уже нельзя будет воспользоваться простой линейной регрессией, поскольку нам не известны параметры распределения времени от заражения до развития манифестной формы, и они определяются нелинейным (в разбираемом примере — экспоненциальным) соотношением.

Гипотетический пример динамики развития манифестных форм заболевания

Год от начала эпидемии	Количество манифестных форм
1	8
2	569
3	3 564
4	13 762
5	38 120
6	80 713
7	138 098
8	201 838
9	264 807
10	323 782

Вместе с тем практически все статистические системы, и SAS здесь не является исключением, поддерживают возможность анализа данных методом нелинейной регрессии. Для подобного анализа системе надо сообщить форму уравнения, а она уже сама определит значения неизвестных параметров. Однако для выполнения анализа необходимо описать, как меняется численность зараженных год от года. Поэтому вначале следует создать файл с множителями для исходного числа зараженных:

```

1. DATA coeff;
2. ARRAY nn{10} nn1–nn10;
3. a1=2.5; a2=0.2;
4. DO year=1 TO 10;
5.     DO j=1 TO 10;
6.         IF year–j+1>0 THEN nn[j]=exp(a1*j–a2*j**2);
7.         ELSE nn[j]=0;
8.     IF j=1 THEN nn[j]=1;
9.     END;
10.    OUTPUT;
11. END;
```

Этот программный код создает таблицу коэффициентов, в которых в первом столбце стоят единицы, а в остальных столбцах стоят множители, соответствующие тому, насколько надо умножить исходную численность инфицированных, чтобы получить количество инфицированных, которое заразилось в год, соответствующий порядковому номеру столбца (например, количество инфицированных во второй год будет находиться во втором столбце). При этом, если номер строки меньше номера столбца, то соответствующее значение обращается в ноль. В таблице для анализа строка означает год с момента начала эпидемии, а столбец — количество лиц, у которых могло развиваться заболевание в этот год. Понятно, что если номер строки меньше номера столбца, то это значит, что данные пациенты еще не заразились и, соответственно, не могут влиять на число манифестных форм заболевания в этом году.

После формирования таблицы коэффициентов следует создать файл с наблюдаемым количеством случаев манифестной формы заболевания и объединить его с файлом коэффициентов:

1. DATA cases;
2. INPUT year cases;
3. CARDS;
4. 1 8
5. 2 569
6. 3 3564
7. 4 13762
8. 5 38120
9. 6 80713
10. 7 138098
11. 8 201838
12. 9 264807
13. 10 323782
14. ;
15. RUN;
16. PROC SORT DATA=cases; BY year; RUN;
17. PROC SORT DATA=coeff; BY year; RUN;
18. DATA new;
19. MERGE cases coeff;
20. BY year;
21. RUN;

Подход аналогичен использованному ранее при простейшем анализе с помощью методики обратного расчета.

Далее наступает самый ответственный этап — проведение нелинейного моделирования. Для этого необходимо вызвать процедуру NLIN и передать ей описание уравнения, которое мы пытаемся оценить:

1. PROC NLIN;
2. PARMs N=8 k=5;
3. MODEL cases=N*(nn1*(1-exp(-year/k)) +
4. nn2*(1-exp(-(year-1)/k))+nn3*(1-exp(-(year-2)/k))+
5. nn4*(1-exp(-(year-3)/k))+nn5*(1-exp(-(year-4)/k))+
6. nn6*(1-exp(-(year-5)/k))+nn7*(1-exp(-(year-6)/k))+
7. nn8*(1-exp(-(year-7)/k))+nn9*(1-exp(-(year-8)/k))+
8. nn10*(1-exp(-(year-9)/k)));
9. RUN; QUIT;

Анализируемое уравнение описано при помощи команды MODEL. В нем количество наблюдаемых случаев манифестной формы заболевания (cases) анализируется в зависимости от произведения исходного числа случаев (N) на сумму манифестных форм, являющихся следствием заражения в каждый год. Коэффициенты nn1—nn10 были рассчитаны ранее и представляют собой множители, которые показывают, во сколько раз численность зараженных в данный год

(от 1 до 10) больше исходного числа инфицированных. Выражение $(1 - \exp(-\text{year}/k))$ читателю уже знакомо и описывает то, какой процент из числа инфицированных имел манифестную форму заболевания в данном году. Особенности создания таблицы коэффициентов привели к тому, что если год предшествует году, когда должны были заразиться те или иные пациенты, то соответствующие коэффициенты p_n равны нулю (например, для первого года $p_1 = 1$, а все остальные коэффициенты $p_2 - p_{10}$ равны нулю). Обязательная команда PARMs предоставляет стартовые значения для интересующих исследователя параметров. Поскольку нелинейная регрессия использует итерационный метод, то она старается найти наиболее правильное решение недалеко от стартовых значений. Хотя в большинстве случаев решение будет получено вне зависимости от стартового значения, в некоторых оно может оказаться неоптимальным (т.н. локальный минимум или максимум). Поэтому лучше всего задавать в качестве исходных параметров значения, которые являются наиболее возможными (например, исходное число, равное количеству манифестных случаев заболевания в первый год эпидемии, а k — пропорционально медианной продолжительности инкубационного периода заболевания из литературы).

Результаты анализа с помощью процедуры нелинейной регрессии представлены ниже:

Параметр	Оценка	Процедура NLIN		
		Приближение стандартной ошибки	Приближенные 95% доверительные пределы	
N	100,0	0,00187	99,9965	100,0
k	11,5417	0,000262	11,5411	11,5423

Видно, что система точно определила исходное число инфицированных (100 человек), а также рассчитала значение коэффициента пропорциональности для экспоненциального уравнения. Поскольку $k = -\frac{t}{\ln(0,5)}$, медианный инкубационный период равен $t = -k * \ln(0,5) = 0,693 * k$. Соответственно, в данном примере медианный инкубационный период равен $11,5417 * 0,693 = 8$ годам, т.е. он также рассчитан точно.

Естественно, что процесс развития манифестной формы заболевания может подчиняться иной форме распределения времен инкубационного периода. Очень часто его описывают при помощи распределения Вейбулла. Распределение Вейбулла описывается следующей формулой:

$$y(t) = 1 - e^{-\left(\frac{t-\theta}{\sigma}\right)^c},$$

где θ называется пороговым значением, σ — нормирующим параметром, а c — параметром формы.

Очевидно, что в случае, если пороговое значение θ равно нулю, а параметр формы c равен единице, то распределение Вейбулла превращается в экспоненциальное распределение с нормирующим параметром σ , равным k .

Пороговое значение позволяет учесть некоторую «отсрочку» в наступлении эффекта. Чем выше пороговое значение, тем позже начнется массовое развитие манифестных форм заболевания. С другой стороны, параметр формы описывает скорость развития манифестных форм после того, как «процесс пошел». Чем он выше, тем быстрее зараженные лица перейдут в состояние с клинически выявляемыми формами заболевания.

Во многих прикладных задачах пороговое значение принимают равным нулю, и тогда значения распределения Вейбулла начинают зависеть только от двух параметров — нормирующего параметра σ и параметра формы c ¹³.

Соответственно, если теперь попытаться провести нелинейное моделирование с использованием вместо экспоненциального распределения времен до развития манифестной формы распределения Вейбулла, то надо лишь немного изменить формулу:

```

1.  PROC NLIN DATA=new;
2.  PARMs N=8 k=5 c=2;
3.  MODEL cases=N*(
4.  nn1*CDF('WEIBULL',year,c,k)+
5.  nn2*CDF('WEIBULL',year-1,c,k)+
6.  nn3*CDF('WEIBULL',year-2,c,k)+
7.  nn4*CDF('WEIBULL',year-3,c,k)+
8.  nn5*CDF('WEIBULL',year-4,c,k)+
9.  nn6*CDF('WEIBULL',year-5,c,k)+
10. nn7*CDF('WEIBULL',year-6,c,k)+
11. nn8*CDF('WEIBULL',year-7,c,k)+
12. nn9*CDF('WEIBULL',year-8,c,k)+
13. nn10*CDF('WEIBULL',year-9,c,k)
14. );
15. RUN; QUIT;

```

В данном программном коде использована встроенная функция SAS по расчету распределения (кумулятивной функции распределения — CDF) для распределения Вейбулла (ключевое слово *Weibull*). Функция возвращает значения количества лиц, у которых в данном году разовьется манифестная форма заболевания в зависимости от года (*year*), параметра формы (*c*) и нормирующего параметра, который мы для сравнимости с предыдущим кодом также обозначили буквой *k*. Если запустить программу нелинейной регрессии на данных предыдущего примера, то будет получен следующий результат:

Параметр	Оценка	Приближение стандартной ошибки	Приближенные 95% доверительные пределы	
N	99,9957	0,0140	99,9627	100,0
k	11,5408	0,00229	11,5354	11,5462
c	1,0000	0,000027	0,9999	1,0001

Видно, что результаты аналогичны моделированию с использованием экспоненциального распределения и система установила, что параметр формы равен 1 (что переводит распределение Вейбулла в экспоненциальное распределение).

После того, как на гипотетических примерах становится понятно, как работает система обратных расчетов, следует попытаться проанализировать реальные данные. Для примера возьмем сведения о регистрации ВИЧ-инфекции и случаях СПИД на территории Российской Федерации [7].

¹³ В литературе их иногда обозначают β и α , соответственно.

Для начала следует описать траекторию развития эпидемии в Российской Федерации. По данным регистрации случаев ВИЧ-инфекции с 1992 по 2003 гг. построим квадратично-экспоненциальное уравнение.

```
1. DATA rf_hiv;
2. INPUT cases;
3. year=_n_;
4. CARDS;
5. 726
6. 163
7. 197
8. 1515
9. 4358
10. 4057
11. 19991
12. 59275
13. 88577
14. 52349
15. 39699
16. 37336
17. ;
18. RUN;
19. PROC NLIN;
20. PARMS N=600 a1=3 a2=0.5;
21. MODEL cases=N*exp(a1*year-a2*year**2);
22. RUN;
```

Данная программа считывает данные по количеству зарегистрированных случаев ВИЧ в Российской Федерации (обратите внимание на то, что года присваиваются автоматически при помощи внутреннего счетчика `_n_`, который принимает значения номера строки, на которой расположена только что введенная переменная), затем вызывается программа нелинейной регрессии, которая пытается найти параметры a_1 и a_2 для квадратично-экспоненциального уравнения.

В результате анализа удается установить, что параметр $a_1 = 3,4343$, а $a_2 = 0,1849$. Теперь можно использовать описанный выше программный код для анализа параметров распределения Вейбулла, создав файл коэффициентов на основании параметров квадратично-экспоненциального уравнения, выявленных на предшествующем этапе, откорректировав их на тот факт, что количество лиц наблюдения увеличилось:

```
1. DATA cases;
2. INPUT cases;
3. year=_n_;
4. CARDS;
5. 30
```

```

6. 44
7. 79
8. 117
9. 167
10. 238
11. 304
12. 344
13. 366
14. 517
15. 720
16. 1230
17. ;
18. RUN;
19. DATA coeff;
20. ARRAY nn{12} nn1–nn12;
21. a1=3.4343; a2=0.1849; a0=–4.7196;
22. DO year=1 to 12;
23.     DO j=1 to 12;
24.         IF year–j+1>0 THEN nn[j]=exp(a1*j–a2*j**2+a0);
25.         ELSE nn[j]=0;
26.         IF j=1 THEN nn[j]=1;
27.     END;
28.     OUTPUT;
29. END;
30. PROC SORT DATA=cases; BY year; RUN;
31. PROC SORT DATA=coeff; BY year; RUN;
32. DATA new;
33.     MERGE cases coeff;
34.     BY year;
35. RUN;
36. PROC NLIN DATA=new;
37. PARS N=1 k=14.5 c=0.93;
38. MODEL cases=N*(
39. nn1*CDF('WEIBULL',year,c,k)+
40. nn2*CDF('WEIBULL',year–1,c,k)+
41. nn3*CDF('WEIBULL',year–2,c,k)+
42. nn4*CDF('WEIBULL',year–3,c,k)+
43. nn5*CDF('WEIBULL',year–4,c,k)+
44. nn6*CDF('WEIBULL',year–5,c,k)+
45. nn7*CDF('WEIBULL',year–6,c,k)+
46. nn8*CDF('WEIBULL',year–7,c,k)+
47. nn9*CDF('WEIBULL',year–8,c,k)+
48. nn10*CDF('WEIBULL',year–9,c,k)+

```

```

49. nn11*CDF('WEIBULL',year-10,c,k)+
50. nn12*CDF('WEIBULL',year-11,c,k)
51. );
52. RUN; QUIT;

```

В данном коде следует обратить внимание на стартовые показатели (команда PARAMS). Как уже отмечалось выше, правильное указание стартовых параметров является крайне важным, особенно в случаях, аналогичных данному, когда имеется достаточно большое количество локальных минимумов и определить правильные значения параметров сложно¹⁴. Результаты приведены ниже:

Параметр	Оценка	Приближение стандартной ошибки	Приближенные 95% доверительные пределы	
N	0,0152	0,00135	0,0122	0,0181
k	14,5000	.	.	.
c	0,9300	.	.	.

Анализ показателей распределения Вейбулла (k и c) показывает, что медианный срок до наступления манифестной формы заболевания (СПИД) составляет чуть менее 10 лет (9 лет и 9 месяцев), что не противоречит известным данным о течении ВИЧ-инфекции. Кроме того, параметр c близок к единице, поэтому без большой ошибки можно считать, что процесс перехода ВИЧ-инфекции в СПИД может быть описан в российской популяции простым экспоненциальным уравнением (с одним параметром), а не распределением Вейбулла (этот вывод будет важен для динамического моделирования, базирующегося на экспоненциальном распределении). Вместе с тем, если попытаться оценить, какое количество случаев ВИЧ-инфекции лежат в основе наблюдаемого количества случаев СПИД, то окажется, что их должно было бы быть всего лишь 5022, а никак не более 270 тыс., которые были официально зарегистрированы в РФ на 2003 год. Если же взять официальные данные по регистрации и проанализировать их с использованием найденного уравнения распределения времени до развития СПИД, то окажется, что в РФ должно было быть отмечено уже более 73 тыс. случаев СПИД, а не 1200, как указывалось ранее. Расхождение этих данных может указывать на три возможных объяснения — либо ВИЧ-инфекция в РФ протекает достаточно доброкачественно (и даже 10 летний инкубационный период является завышенным), либо существует значительная недорегистрация случаев СПИД, либо наиболее адекватным описанием процесса трансформации ВИЧ-инфекции в СПИД является какое-то более сложное распределение, а не распределение Вейбулла или экспоненциальное.

Таким образом, методы обратного расчета, несмотря на кажущуюся простоту и логичность, обладают рядом серьезных недостатков, часть из которых является следствием необходимости описать процесс течения эпидемии и перехода одной формы в другую относительно простой зависимостью. Вторая техническая трудность сопряжена с тем, что оценка нелинейной регрессии не вызывает сложностей только в простейшем случае. На реальных данных результат может оказаться неопределенным или требовать значительных усилий по поиску правильного решения. Кроме того, методы обратного расчета страдают от всех дефектов экстраполяционных методик, и они были разработаны только для одного заболевания (ВИЧ-инфекции, хотя аналогичный подход можно было бы использовать для туберкулеза), для которого диагностика была возможна только на поздних стадиях.

Эти проблемы привели к росту интереса к динамическому моделированию.

¹⁴ На самом деле в данном случае использовалась другая процедура — NLMIXED, которая позволяет оценить уравнение при помощи метода максимального правдоподобия. Затем уже найденные стартовые параметры были использованы в процедуре NLIN.

3. Динамическое моделирование

Динамическое моделирование заключается в том, что аналитик пытается описать математическими закономерностями процесс распространения заразного процесса в популяции. Динамическое моделирование выступает в двух основных ипостасях — детерминистского и стохастического. Детерминистское опирается на свойства больших групп и базируется на использовании математического аппарата, описывающего непрерывные процессы. Стохастическое пытается моделировать процесс аналогично тому, как это происходит в реальности — моделируются взаимодействия между отдельными объектами (например, людьми) и, если один из этих объектов заражен, то второй может с определенной вероятностью заразиться. В таких моделях вопрос о том, окажется или нет «зараженным» второй субъект, определяется генератором случайных чисел. Наперед никто не может сказать, окончится ли контакт заразного и уязвимо заражением, однако генератор случайных чисел конфигурируется так, чтобы для всей популяции в целом количество контактов, в которых произошло «заражение», соответствовало наблюдаемым данным.

Надо заметить, что стохастическое моделирование более «реально» описывает распространение инфекционного заболевания и может быть легко адаптировано к ситуации, когда население достаточно гетерогенно по числу контактов, однако оно требует моделирования каждого контакта и поэтому не очень просто для изучения распространения заразного процесса в больших популяциях. В больших популяциях допущения, используемые при формулировке детерминистских моделей, оказывают меньшее влияние на результат, и поэтому для целей моделирования распространения в больших популяциях основным инструментом продолжает оставаться детерминистское моделирование.

3.1. Детерминистские модели

3.1.1. Создание графов динамики заболевания

Детерминистское моделирование начинается с того, что создается граф возможного развития заболевания. Например, если заболевание передается воздушно-капельным путем в большом городе (популяция относительно гомогенна) и не оставляет после себя иммунитета, то вся популяция будет разделяться на две группы — инфицированных (заразных) и уязвимых (переболевших или никогда не болевших). Лица, переболевшие заболеванием, не имеют иммунитета перед последующим заражением, и поэтому они возвращаются в группу уязвимых. В моделировании инфекционных заболеваний подобные модели называются SIS (от англ. 'Susceptible-Infected-Susceptible' или Уязвимый-Инфицированный-Уязвимый).

Графы, соответствующие каждой модели, лучше всего создавать в программах динамического моделирования, таких как Vensim или Stella.

Для начинающих рекомендуется познакомиться с одной из этих систем, например, с системой Vensim (Ventana Systems, Inc.), которая доступна для бесплатного скачивания и использования в учебных целях (конкурирующая система, Stella, которая немного проще в работе, в настоящий момент в форме бесплатной демонстрационной программы отсутствует, однако можно приобрести программу с заблокированной возможностью сохранения созданных моделей за 50 долларов (ранее эта версия предлагалась бесплатно)).

В системе Vensim после установки и запуска появляется пустое окно с панелью выбора элементов для моделирования (рис. 1).



Рис. 1. Панель инструментов системы Vensim

Основными элементами, которые будут использоваться при построении моделей, будут инструменты 2–5. Вначале следует воспользоваться инструментом 3 (Box Variable), который иногда в системах динамического моделирования называется контейнером (в системе Vensim его называют коробчатой переменной). Такое название вызвано тем, что его можно сравнить с контейнером или сосудом для жидкости, откуда содержимое будет перетекать в другие сосуды. Контейнеры обычно на графах обозначаются в виде прямоугольников. Соответственно, поскольку, например, уязвимые «перетекают» в группу «инфицированных», то для этих двух групп надо будет воспользоваться контейнерами. После щелчка мышью по этому инструменту следует щелкнуть на белом рабочем поле системы Vensim и в появившемся окошке ввести имя для контейнера. По умолчанию система не очень любит русские буквы, однако, если заменить в настройках шрифт по умолчанию на кириллический (например, Arial Cyr), то можно будет делать подписи и по-русски.

Создав контейнеры для состояний, следует указать, откуда куда что может «перетекать». Это делается при помощи инструмента 5 (Rate). В системе Vensim этот инструмент носит название «скорость», однако мы будем его назвать «поток». Потоки принято обозначать в виде двойной стрелки с изображением, напоминающим вентиль посередине. Щелкнув по этому инструменту на панели инструментов, надо однократно щелкнуть на том контейнере, из которого вытекает поток, а затем однократно щелкнуть на том контейнере, в который поток направляется. В появившемся окне следует ввести имя потока.

Затем можно сделать получившийся граф более красивым, передвинув и изменив размеры контейнеров и потоков при помощи инструмента 1 (Move/Size Words and Arrows). Для этого, щелкнув по инструменту, следует щелкнуть по изменяемому элементу и затем мышью либо потянуть его границы, либо передвинуть.

Вот как выглядит простейшая описанная выше модель инфекционного заболевания в программе Vensim (рис. 2).



Рис. 2. Граф болезненных состояний

Наша система состоит из двух контейнеров (уязвимые и инфицированные) и двух потоков (заражение и выздоровление). Уязвимые заражаются и становятся инфицированными, последние выздоравливают и опять становятся уязвимыми к заражению.

Теперь следует разобраться с тем, какие процессы определяют скорость потока (заражения или выздоровления). Очевидно, что количество новых заражений тем выше, чем больше

в популяции инфицированных. Соответственно, количество инфицированных должно оказывать воздействие на скорость потока заражений. Вторым важнейшим показателем является количество контактов в единицу времени (этот показатель принято обозначать буквой c). Действительно, чем больше контактов происходит между людьми в популяции, тем выше вероятность встречи заразного пациента и уязвимого и передачи инфекции. Вместе с тем не все контакты приводят к заражению — есть более и менее заразные заболевания и соответственно лишь часть контактов окажется контактами, в результате которых произошло заражение. Вероятность заражения при однократном контакте обозначается буквой β . Довольно часто обсуждают не количество контактов и по отдельности вероятность заражения при однократном контакте, а объединяют их в количество контактов, при которых может произойти заражение. Эта величина равна произведению $c * \beta = \lambda$, и определяется как количество контактов заразного человека в единицу времени, в результате которых происходит заражение. Например, если зараженный человек в день контактирует с 10 другими людьми и заразность заболевания составляет 50%, то количество контактов, в результате которых произошло заражение, будет составлять $\lambda = 10 * 50\% / 100\% = 5$. Хотя при таком объединении количество необходимых для отслеживания переменных уменьшается, большого выигрыша это не дает и может оказать негативное влияние при попытках моделировать разные группы с одинаковой вероятностью заражения, но разным количеством контактов. Поэтому в дальнейшем мы будем пользоваться показателями числа контактов и вероятностью заражения при однократном контакте по отдельности.

Итак, скорость заражения в нашем примере определяется числом контактов, вероятностью заражения при однократном контакте и количеством инфицированных. Если задуматься, это вполне очевидно, поскольку количество новых случаев инфекции зависит от того, сколько человек заразит каждый инфицированный, а это, в свою очередь, зависит от числа его контактов с людьми и вероятности передачи инфекции при однократном контакте.

Здесь есть, однако, одна небольшая сложность. Инфицированный человек не ищет (мы надеемся) уязвимых людей для своих контактов, а контактирует со всеми подряд, однако часть этих контактов приходится на инфицированных, а часть — на уязвимых. Поэтому, чем больше в популяции инфицированных (меньше уязвимых), тем большее количество контактов происходит в парах инфицированный—инфицированный, в которых заражение невозможно. Поэтому скорость заражения также зависит от пропорции уязвимых среди всех контактов.

Количество инфицированных увеличивается за счет заражения и уменьшается за счет выздоровления. Скорость выздоровления зависит от того, как долго продолжается заболевание (скорость выздоровления — величина, обратная длительности заболевания). Очевидно, что поток выздоровлений зависит от количества инфицированных и скорости выздоровления.

Итак, для описания модели следует ввести несколько новых параметров (количество контактов, вероятность заражения при однократном контакте, длительность заболевания). Параметры, которые не меняются со временем, в системе Vensim называются вспомогательными переменными или постоянными (Variable — Auxillary/Constant), и для их изображения используется инструмент 2. Для того, чтобы ввести новый параметр, следует щелкнуть по инструменту 2, а затем щелкнуть где-то на рабочем поле и ввести имя нового параметра.

После того, как все параметры были внесены на граф, следует указать, на какой поток они влияют (и какие еще показатели влияют на скорость потока). Делается это при помощи инструмента 4 «стрелка» (Arrow). Для того, чтобы указать, какие показатели (содержимое контейнера, параметры) влияют на скорость потока, надо щелкнуть по инструменту 4, затем по тому показателю, который влияет, а затем по «вентилю» на стрелке, обозначающему поток. Между этими двумя объектами возникнет голубая стрелка. Модель с указанными связями между объектами показана на рис. 3.

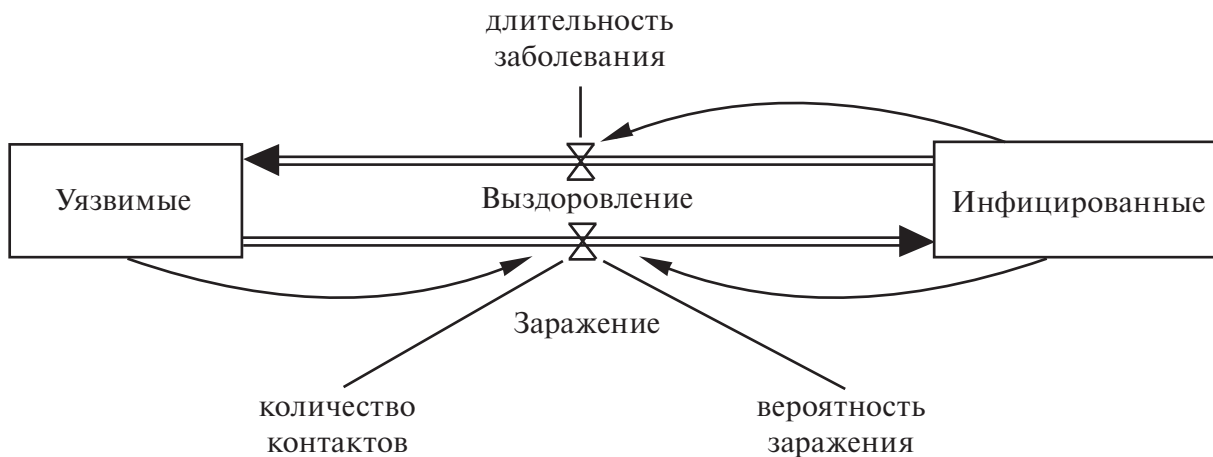


Рис. 3. Граф простейшей модели распространения заразного заболевания (SIS-модель)

3.1.2. SIS-модели

Создание SIS-моделей в системе Vensim. На рис. 3 изображен полный граф простейшей модели распространения заразного заболевания. Однако модель еще не закончена, ибо не указаны значения параметров, начальные условия для моделирования и правила для определения скорости потоков. Вначале следует указать значения для скоростей потоков. Для того, чтобы проще получить доступ ко всей необходимой информации, следует перевести модель в режим редактирования выражений (инструмент 10, Equations). Vensim автоматически подсвечивает те объекты, для которых отсутствует необходимая информация. Щелчок левой кнопкой мыши по объекту, информацию о котором аналитик хочет отредактировать, вызывает окно редактирования формул, показанное на рис. 4.

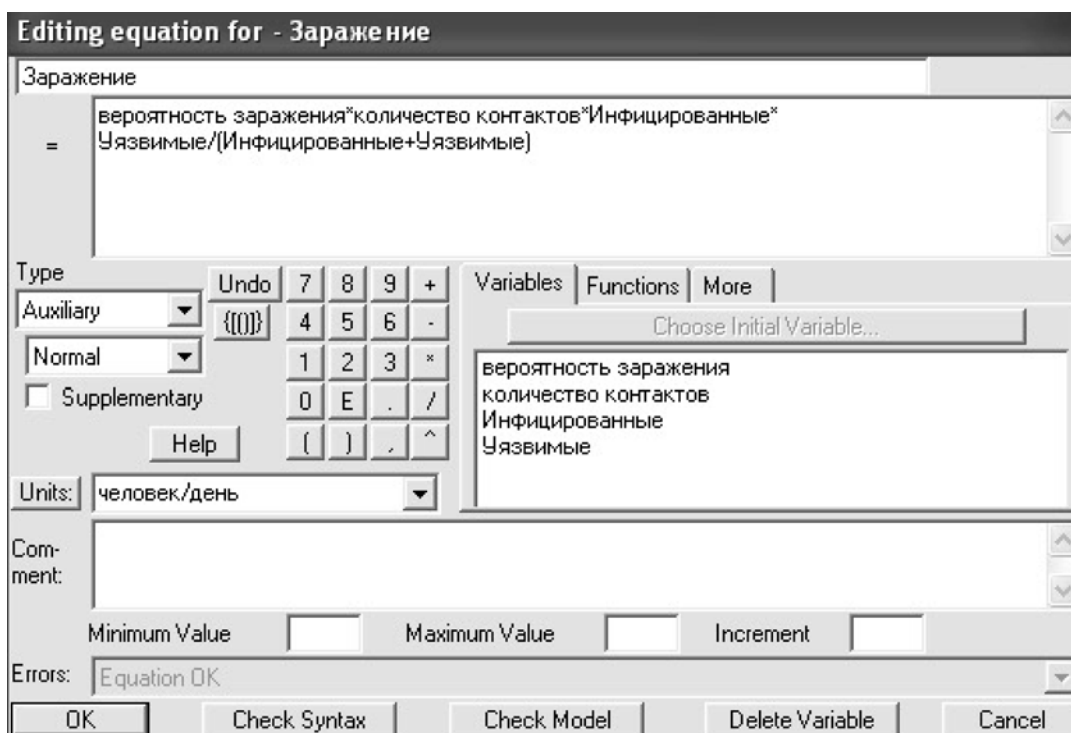


Рис. 4. Окно редактирования формул (анализируется поток «заражение»)

В этом окне аналитик должен ввести выражение, определяющее скорость потока. На рис. 4 те переменные и параметры, которые влияют на скорость потока (содержимое двух контейнеров и два параметра) приведены в окошке с закладкой Variables. Ранее указывалось, что скорость заражения зависит от количества контактов в единицу времени, умноженного на вероятность заражения, умноженного на число заразных людей в популяции и на относительное число уязвимых в популяции. Формула вводится щелчком по имени переменной и по кнопкам цифр и алгебраических операторов на панели рядом с именами переменных. В окошке после знака равенства введена полная формула. Она выглядит так:

$$\text{Заражение} = \text{Вероятность заражения} * \text{Количество контактов} * \text{Инфицированные} * [\text{Уязвимые} / (\text{Уязвимые} + \text{Инфицированные})]$$

Переменные, входящие в модель, могут быть модифицированы при помощи функций, содержащихся на закладке Functions, а также для формул могут быть использованы логические правила с закладки More.

В нашем простейшем примере следует ограничиться только немодифицированными значениями переменных.

После нажатия на кнопку ОК формула сохраняется и подсветка данного объекта снимается, так что аналитик видит, какие еще объекты следует отредактировать.

Для скорости выздоровления вводимая формула будет:

$$\text{Выздоровление} = \text{Инфицированные} / \text{длительность заболевания}$$

При определении формул для контейнеров система Vensim открывает окно с автоматически составленным уравнением, которое равно разности притекающих и оттекающих потоков, например, для контейнера «Инфицированные» она равна:

$$\text{Инфицированные} = \text{Заражение} - \text{Выздоровление}$$

Соответственно, для контейнера «Уязвимые» она равна:

$$\text{Уязвимые} = -\text{Заражение} + \text{Выздоровление}$$

Следует обратить внимание на то, что в записи скорость заражения стоит первой с отрицательным знаком. Конечно, можно использовать и более стандартную запись $\text{Выздоровление} - \text{Заражение}$, однако следует помнить, что выздоровление не является обязательным (этот класс моделей будет разобран позднее), и тогда формула приобретет несколько странный вид: $\text{Уязвимые} = -\text{Заражение}$.

Все описанные выше формулы для контейнеров требуют еще одного условия — начального. При составлении модели следует указать, сколько единиц (человек), находится в каждом состоянии (контейнере) к началу моделирования. Если помнить, что у каждого контейнера есть начальная «наполненность», тогда становится понятным, что правильная трактовка формулы для контейнера означает: численность объектов в контейнере равна численности объектов в предыдущем временном периоде минус количество объектов, ушедших с исходящими потоками и пришедших — с входящими.

На самом деле такое суммирование небольших изменений математически определяется как интегрирование и, соответственно, скорости истекающего и приходящего потоков будут определяться дифференциальными уравнениями. Это отступление окажется важным для дальнейшего обсуждения моделирования в системе SAS.

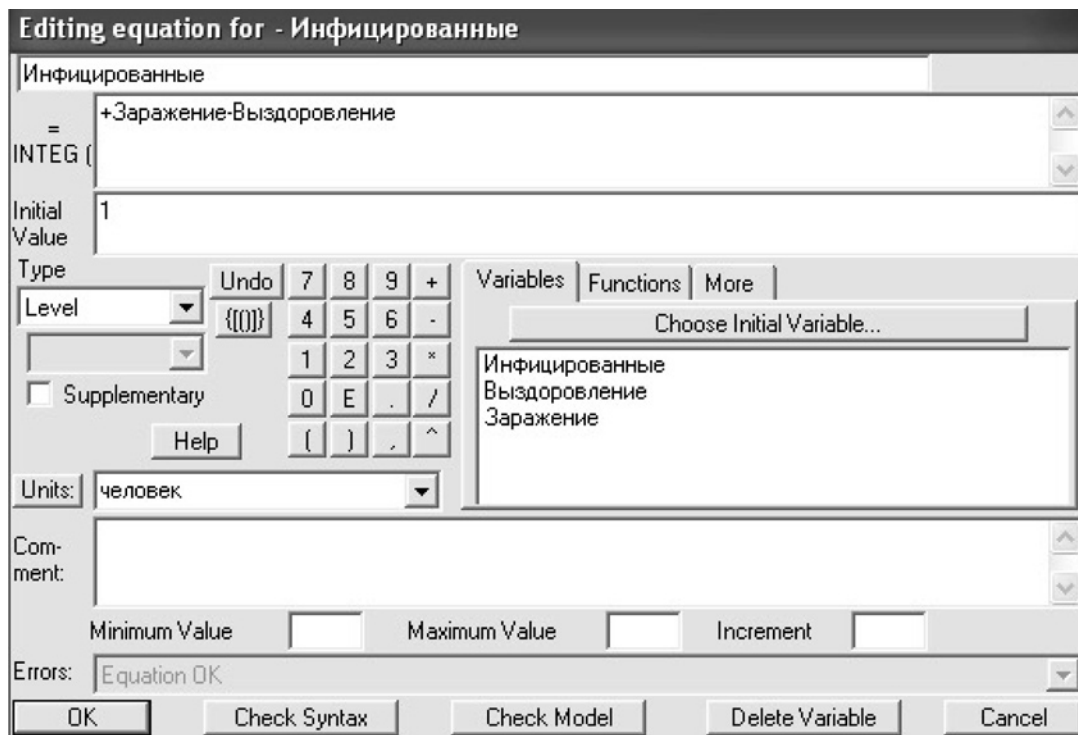


Рис. 5. Окно редактирования формул (анализируется контейнер «инфицированные»)

Итак, после указания формулы для контейнера обязательно надо указать начальные условия, так, как это показано на рис. 5.

Если сравнить рис. 4 и рис. 5, то видно, что на рис. 5 появилось дополнительное окошко для начального значения (Initial Value), которое равно 1, а в самой строке формулы указано, что «наполненность» контейнера является следствием интегрирования двух потоков — в случае, показанном на рис. 5 — потока заражений и потока выздоровлений. Для контейнера «Уязвимые» установим начальное значение равным 10 000.

Следующим этапом будет установление значений параметров. В данном случае это количество контактов, вероятность заражения и длительность заболевания.

Окно редактирования формул для параметров не отличается от такового для потоков, однако, если на параметр не влияют другие параметры, то окошко переменных будет пустым и аналитику надо будет вставить в окно значения постоянную величину. Будем считать, что существующие данные показывают, что количество контактов составляет 5 в день, вероятность заражения при однократном контакте — 0,1 (10% всех контактов), а длительность заболевания 10 дней. На рис. 6 показано это окно для параметра «количество контактов», где количество контактов равно 5 в день.

В реальных моделях параметр может зависеть от других параметров или даже от наполненности контейнеров. Кроме того, в крупных проектах рекомендуется указывать единицы, в которых измеряется тот или иной параметр (окошко units).

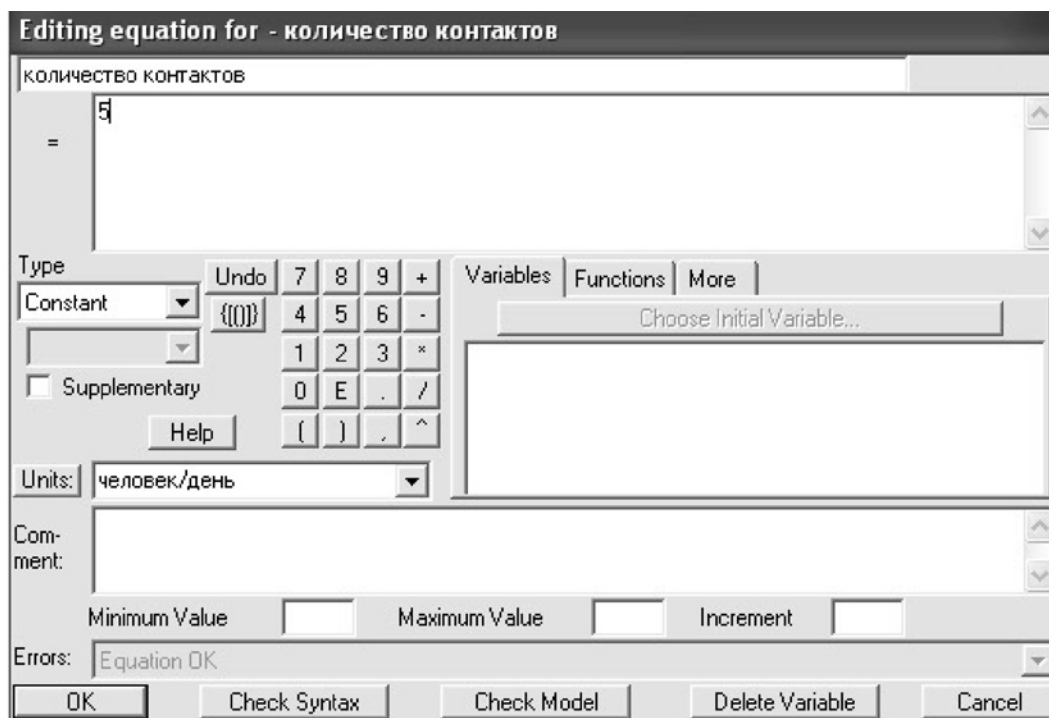


Рис. 6. Окно редактирования формул (представлен параметр «количество контактов»)

В небольшом проекте это не очень важно, однако в большом может помочь найти неправильные формулы или недоучтенные параметры.

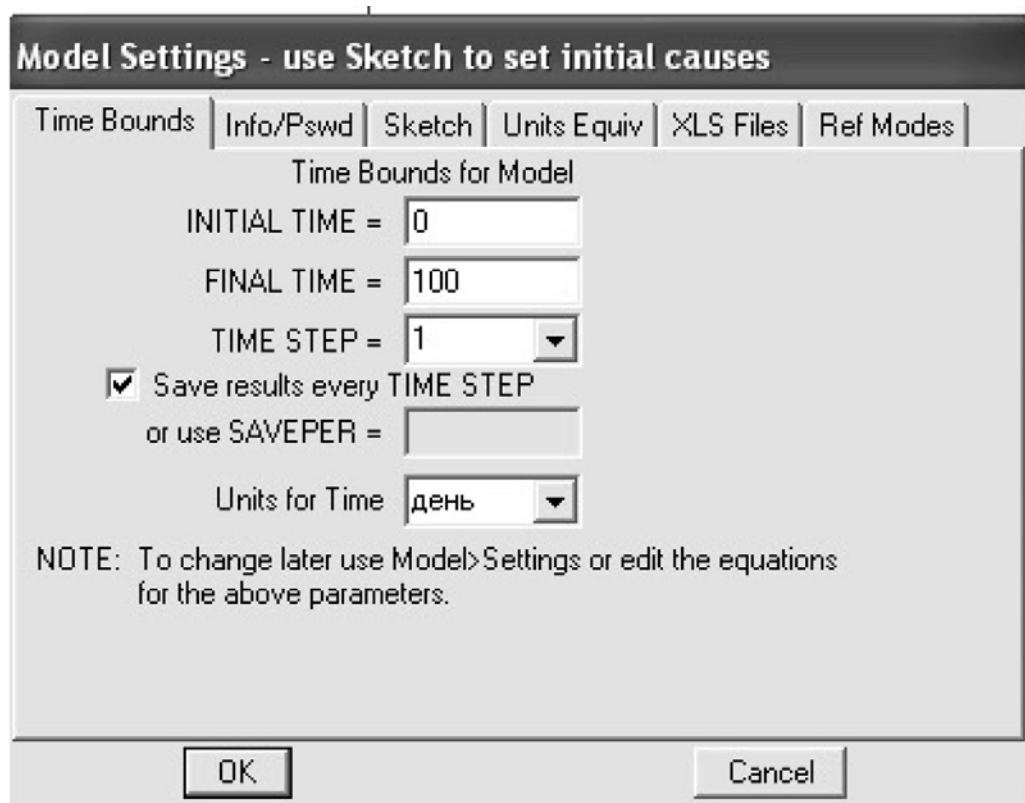


Рис. 7. Закладка настроек времени моделирования

Прежде чем начинать работать с единицами измерения, следует установить базовые единицы времени, с которыми будет работать данный проект. Предположим, что в данном случае базовой единицей времени будет один день. Для того, чтобы сообщить системе о нашем решении, следует в меню Модель (Model) выбрать пункт Настройки (Settings) и в появившемся окошке выбрать закладку Временные пределы (Time Bounds). На этой закладке, показанной на рис. 7, аналитик устанавливает временные пределы для моделирования (например, от 0 до 100 дней после заноса инфекции) и указывает единицы измерения времени. По умолчанию Vensim располагает предустановленными единицами для дня, месяца, года, квартала, часа и секунды, однако названия единиц времени идут по-английски. Поэтому, кликнув мышью по окошку Единицы времени (Units for Time), можно по-русски вписать единицу измерения «день», как показано на рис. 7.

После этого надо подумать, не будет ли у нас среди единиц измерения синонимов. Например, контакт и человек являются синонимами, и об этом надо сообщить системе. Делается это на закладке Эквивалентность единиц (Units Equiv), того же окна настроек (см. рис. 8). Синонимы перечисляются через запятую в нижней строке ввода этой закладки, а затем добавляются нажатием на кнопку Добавить отредактированное (Add Editing).

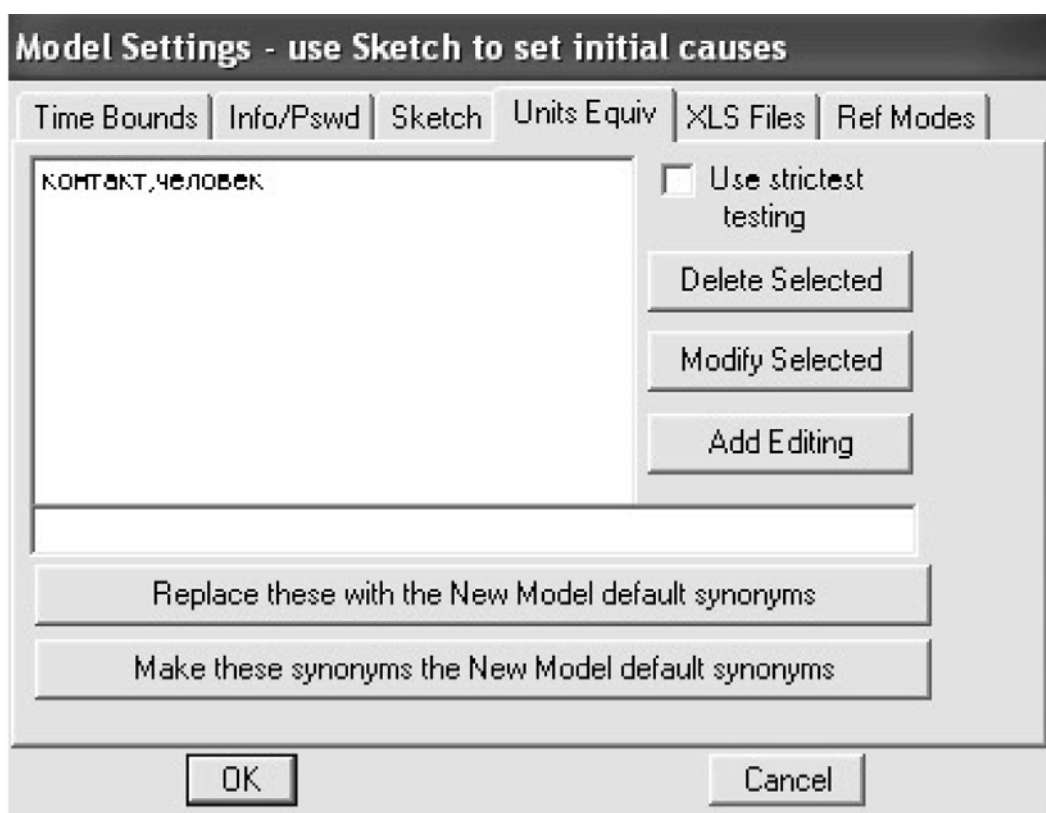


Рис. 8. Закладка настроек эквивалентности единиц измерения

После этой операции следует нажать кнопку ОК, и все новые настройки будут сохранены. Теперь можно переходить к определению единиц для каждого объекта. Делается это, как уже указывалось, при помощи все тех же окошек редактирования формул, путем ввода единиц в строку Единицы (Units). Если единицы уже описаны (были введены ранее), то их список можно посмотреть, щелкнув по стрелочке в конце строки Units.

Для количества контактов единицей будет «человек/день», для вероятности заражения «1/контакт» (заражений на 1 контакт), для времени длительности заболевания — «день», для уязвимых и инфицированных — «человек». Поток измеряется в «человек/день». Для

редактирования формул надо щелкнуть правой кнопкой мыши по соответствующему объекту и выбрать в появившемся окошке кнопку Уравнение (Equation).

После ввода всех единиц следует провести проверку размерности модели которую можно вызвать в меню Модель (Model) пунктом Проверка размерности (Units Check) или просто нажатием комбинации клавиш Ctrl-T.

Если все формулы имеют правильную размерность, система сообщит об этом, в противном случае появится окошко, которое будет содержать описание формулы, где найдена ошибка, и указано, какая размерность с правой и с левой стороны уравнения не совпадают.

Указание единиц нужно только для проверки размерности, и поэтому, как уже указывалось выше, не является обязательным, однако работа по вводу единиц воздается сторицей при анализе сложных моделей и документировании результатов.

После завершения анализа размерности модель готова к использованию, и можно запускать ее анализ. Для этого надо всего лишь нажать на кнопку запуска анализа (Run a Simulation) или нажать сочетание клавиш Ctrl-R. Система поинтересуется, можно ли записать новые данные поверх старых (всегда используется одна и та же рабочая область) и, если анализ завершится без каких-либо сообщений, то завершение было благополучным, без каких-либо ошибок. Теперь результаты анализа можно посмотреть в графическом или табличном виде. Учебная версия Vensim позволяет рисовать графики только по одному на каждый объект.

Поэтому для составления графика следует щелкнуть по объекту, который интересует аналитика, и нажать на кнопку График (Graph), расположенную на левой стороне рабочей области (рис. 9).



Рис. 9. Кнопки анализа модели (верхняя — график, нижняя — таблица)

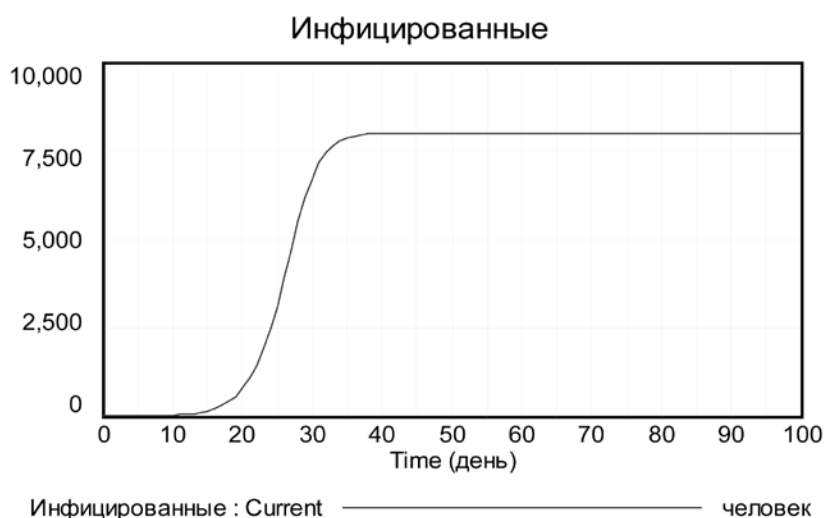


Рис. 10. Результаты анализа гипотетической модели распространения заразного заболевания

После нажатия на кнопку появляется график, аналогичный графику, представленному на рис. 10 для количества инфицированных. Видно, что количество инфицированных быстро возрастает между 15 и 30 днями эпидемии, а затем стабилизируется на уровне немногим более 8000 человек. Анализ данных по потокам зараженных и выздоравливающих показывает, что после короткого периода он стабилизируется на уровне 800 человек в день.

Далее можно изменять входные параметры модели и анализировать влияние этих изменений на течение моделируемого эпидемического процесса.

Анализ простейших SIS-моделей в SAS. В принципе весь анализ можно делать в системе Vensim, однако в бесплатной версии возможности ограничены, и для более полных возможностей по анализу модели придется воспользоваться другими системами¹⁵. При этом есть возможность воспользоваться профессиональными системами статистической обработки данных, такими, например, как SAS.

К сожалению, в SAS нет столь удобного графического интерфейса для описания модели, поэтому начинающим аналитикам стоит порекомендовать вначале построить модель, или по крайней мере ее графическое представление, в системе Vensim, а затем транслировать ее в систему SAS. Когда аналитик по привыкнет к использованию системы SAS после Vensim, он сможет формулировать описание модели напрямую на языке SAS, однако даже в этом случае разумным является иметь перед глазами графическое изображение модели, подобное тому, что представлено на рис. 3.

В SAS для моделирования заразного процесса лучше всего использовать процедуру MODEL, которая поддерживает решение различных задач математического моделирования, в том числе решение систем дифференциальных уравнений, которые стоят за всеми детерминистскими моделями. Для формулировки модели в SAS можно воспользоваться графом модели, представленным на рис. 3. Формулы для потоков вводятся так же, как это делалось в системе Vensim (однако использование русских букв в именах переменных не допускается), а вот для контейнеров необходимо указывать, что содержимое контейнера накапливается путем суммирования потоков (перед именем контейнера добавляется приставка DERT., означающая, что для получения значения содержимого контейнера приведенное выражение надо интегрировать ('dert' — сокращение от английского 'derivative' — производная)).

Если обозначить число контактов буквой «с», вероятность заражения словом «beta», а длительность заболевания буквой «D», то выражения для потоков заражения (Infected) и выздоровления (Cured) будут выглядеть так (принимая во внимание, что мы обозначаем численность инфицированных на данный момент времени как Inf, а число уязвимых — Sus):

$$\text{Infected} = c * \text{beta} * \text{Inf} * \text{Sus} / (\text{Inf} + \text{Sus})$$

Выражение для потока выздоравливающих тогда будет выглядеть так:

$$\text{Cured} = \text{Inf} / D$$

Соответственно, выражение для численности инфицированных (наполнение контейнера) задается выражением:

$$\text{DERT.Inf} = \text{Infected} - \text{Cured}$$

¹⁵ Или приобрести полную версию Vensim.

Выражение для численности уязвимых задается практически таким же выражением:

$$\text{DERT.Sus} = \text{Cured} - \text{Infected}$$

Эти четыре выражения полностью описывают простейшую модель распространения инфекции в популяции в рамках концепции SIS.

Однако для анализа следует установить еще несколько показателей. Во-первых, следует описать, на протяжении какого периода времени необходимо делать анализ и каков минимальный шаг переменной времени. Делается это путем создания специального файла, который, как минимум, должен содержать одну переменную – time (время). Программа для создания этого файла выглядит так:

1. DATA t;
2. DO time=1 TO 100;
3. OUTPUT;
4. END;
5. RUN;

Здесь создается файл с именем t, в котором есть одна переменная time, принимающая значения от 1 до 100. Файл легко модифицировать, не меняя модели, для удлинения или укорочения времени анализа. Отметим также, что в этом файле могут храниться время-зависимые параметры (например, если со временем количество контактов снижается). Когда файл времени анализа создан, можно начать конфигурировать процедуру моделирования (PROC MODEL) для выполнения анализа.

Как известно из предыдущего обсуждения, пока еще отсутствуют значения у параметров (количество контактов, вероятность заражения, длительность заболевания). Они вводятся в модель либо путем указания значений в команде PARMs (параметры), либо путем создания файла с описанием параметров (последний случай очень удобен, если параметры являются расчетными). В данном случае мы воспользуемся прямым вводом значений параметров, и тогда соответствующая строка кода будет выглядеть так:

PARMS c 5 beta 0,1 D 10;

Здесь после каждого параметра идет его значение (надо заметить, что в случае ввода параметров через файл строка PARMs все равно должна присутствовать, но в ней будут приведены лишь имена этих параметров).

Теперь надо сообщить процедуре моделирования, какие переменные описывают состояния (контейнеры) и их первоначальные значения. Таких переменных в данном примере две, и они описываются командой DEPENDENT (зависимые). Как и в случае команды PARMs, после имени переменной-контейнера идет ее начальное значение.

DEPENDENT Inf1 Sus 10 000;

Как и при моделировании в системе Vensim, мы начинаем модель с одного зараженного, оказавшегося в популяции 10 000 уязвимых. Теперь можно уже сформулировать полное описание модели в системе SAS:

1. PROC MODEL DATA=t;
2. DEPENDENT Inf 1 Sus 10000;
3. PARMS c 5 beta 0.1 D 10;
4. Infected=c*beta*Inf*Sus/(Inf+Sus);
5. Cured=Inf/D;
6. DERT.Inf=Infected—Cured;
7. DERT.Sus=—Infected+Cured;
8. SOLVE Inf Sus/ OUT=determ;
9. RUN; QUIT;

В первой строке рядом с наименованием процедур находится указание на то, как называется файл с описанием временных рамок анализа. Далее идут команды DEPENDENT, описывающая основные переменные анализа, и PARMS, служащая для описания параметров модели.

Следующие четыре строки описывают собственно модель. Команда SOLVE (решить) запускает анализ модели. В качестве параметров команде SOLVE передаются имена переменных, значения которых мы хотим вычислить (число инфицированных и уязвимых), а затем указываются опции (после наклонной черты). В данном случае указывается, что результаты анализа следует сохранить в файле с именем determ. Это необходимо сделать, поскольку результаты анализа будут использоваться для построения графиков и вообще анализа результатов запуска модели.

Для построения графика воспользуемся следующими командами:

1. PROC GPLOT DATA=DETERM;
2. SYMBOL1 V=none I=j w=3;
3. AXIS1 LABEL=('время (дни)');
4. AXIS2 LABEL=(A=90 'количество человек');
5. PLOT (Sus Inf)*time/OVERLAY HAXIS=AXIS1 VAXIS=AXIS2;
6. RUN;

Это обычные команды построения линейного графика, использование которых подробно описано в книге [4]. В результате получается график, показанный на рис. 11.

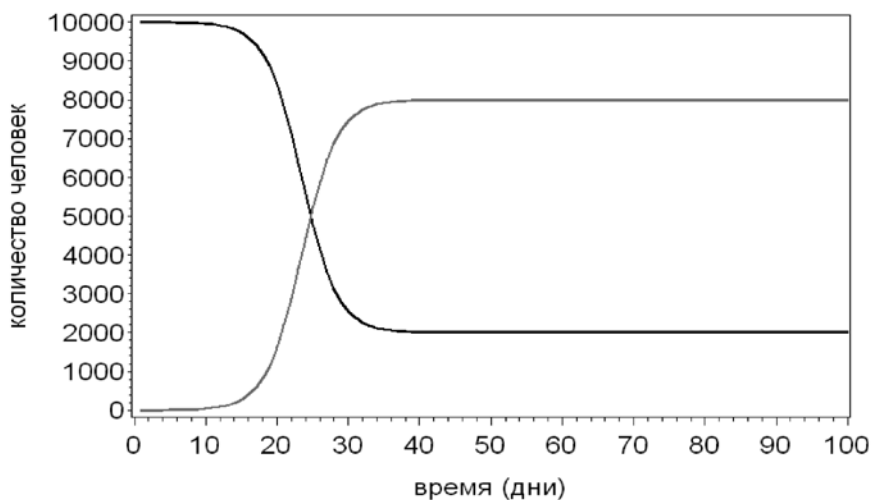


Рис. 11. Результаты анализа SIS-модели в системе SAS

Рис. 11 аналогичен рис. 10, с той лишь разницей, что на один рисунок нанесены одновременно численности и инфицированных, и уязвимых. Аналогичность графиков указывает на правильность построения модели.

SIS-модели и репродуктивное число. При первом же взгляде на рис. 11 и рис. 10 бросается в глаза то, что по прошествии определенного периода времени количество инфицированных и уязвимых стабилизируются. Понятно, что стабильность может наблюдаться тогда, когда количество вновь инфицированных становится равным количеству выздоравливающих (приток равен оттоку).

Если вспомнить выражения для числа инфицированных и выздоравливающих, то легко понять, что условие стабильности выполняется тогда, когда:

$$\frac{I}{D} = \beta * c * I * \frac{S}{S + I},$$

где I — количество инфицированных, D — длительность заболевания, β — вероятность заражения, c — количество контактов и S — число уязвимых.

Слева в этом уравнении стоит число выздоравливающих, справа — число инфицированных.

Достаточно легко продемонстрировать, что это уравнение преобразуется к следующему виду:

$$\frac{S + I}{S} = \beta * c * D$$

Справа, если вдуматься, стоит число, равное тому, сколько человек может заразить один больной за время своей болезни (произведение вероятности заражения на количество контактов — это количество человек, которых он может заразить за единицу времени, а длительность заболевания — сколько таких единиц он заразил). Эта величина называется базовым репродуктивным числом, обозначается R_0 , и играет очень важную роль в понимании динамики распространения инфекций в популяции.

В левой части стоит выражение, обратное проценту уязвимых в популяции. Соответственно, можно записать¹⁶, что:

$$\frac{1}{S} = R_0$$

Отсюда легко вывести, что равновесное количество инфицированных в популяции равно:

$$I = 1 - S = 1 - \frac{1}{R_0}$$

Из этого выражения следует, что в случае SIS-моделей, зная базовое репродуктивное число, мы можем легко рассчитать уровень постоянной заболеваемости заразным заболеванием (уровень эндемии), или, с другой стороны, зная уровень эндемии, можем оценить базовое репродуктивное число.

¹⁶ Если принять, что S измеряется в долях от численности популяции в целом.

Базовое репродуктивное число заслуживает большое внимание потому, что, зная его, можно выяснить, что будет происходить после внесения заразного начала в популяцию. Если базовое репродуктивное число меньше единицы, то каждый заболевший в популяции будет в среднем заражать менее одного человека. Соответственно, второе поколение заболевших будет меньше первого, третье — меньше второго и т.д., что означает, что количество заболевших будет непрерывно снижаться и эпидемический процесс быстро пойдет на спад.

Если же базовое репродуктивное число больше единицы, то число случаев будет нарастать — соответственно, при внесении заразного начала в популяцию начнется эпидемия. Если базовое репродуктивное число равно единице, то количество заразившихся точно равно количеству выздоровевших и уровень распространенности остается постоянным (эндемичный уровень).

Однако задумаемся над тем, что произойдет, если в популяции уже есть люди, нечувствительные к заражению. Тогда, как уже обсуждалось выше, ряд контактов выпадает из числа возможных случаев заражения, и только те контакты, где инфицированный встречается уязвимого, способствуют распространению заразного агента. В этом случае базовое репродуктивное число плохо отражает поведение эпидемического процесса, однако можно воспользоваться другим показателем, который называется просто репродуктивным числом (обозначается R). Этот показатель также отражает число лиц, которое будет заражено одним больным на протяжении его болезни, однако не в результате попадания инфекции в полностью уязвимую популяцию, а в результате ее заноса в популяцию, состоящую из как уязвимых, так и резистентных индивидов. В рамках SIS-моделей такими резистентными индивидами будут только больные (в рамках рассматриваемых позднее SIR-моделей резистентными будут переболевшие и выработавшие иммунитет люди).

Соответственно, количество контактов уменьшается на процент людей, находящихся в резистентной группе (в данном случае больных). Тогда репродуктивное число будет не чем иным, как произведением базового репродуктивного числа на количество уязвимых в популяции:

$$R = S * R_0$$

Выше мы видели, что в случае SIS-моделей в конце концов эпидемический процесс стабилизируется так, что количество инфицируемых становится равным количеству выздоравливающих. Это означает, что репродуктивное число становится равным единице. Это условие можно записать так:

$$R = 1 = S * R_0$$

Отсюда

$$S = \frac{1}{R_0},$$

а

$$I = 1 - S = 1 - \frac{1}{R_0}$$

Мы получили то же самое выражение для определения уровня эндемии в SIS-моделях, которое было выведено ранее, однако уже пользуясь правилом приравнивания репродуктивного числа единице.

Использование репродуктивных чисел позволяет нам сформулировать правила борьбы с эпидемиями. Самое простое из них заключается в том, что, чтобы добиться устранения инфекции из данной популяции, надо добиться снижения базового репродуктивного числа до значения менее единицы. Поскольку известно, что базовое репродуктивное число зависит от числа контактов в единицу времени, вероятности заражения и длительности заразного периода, мероприятия должны быть направлены на ограничение количества контактов, снижение вероятности заражения при контактах (путем использования средств защиты или профилактики) и сокращения длительности заразного периода выявлением и лечением заболевших.

С другой стороны, при заданных условиях эпидемия, распространяющаяся по законам, соответствующим SIS-моделям, будет самоограничивающейся. Она будет достигать уровня эндемии, который равен дополнению до единицы величины, обратной базовому репродуктивному числу, и затем стабилизироваться.

Wasserheit и Aral [21] предлагают обобщенную модель распространения инфекций в рамках SIS-моделей (к ним относятся большинство инфекций, передающихся половым путем). Они считают, что вскоре после внесения нового инфекционного агента заболевание быстро распространяется в популяции и достигает того, что они называют «гиперэндемический уровень». Этот уровень определяется особенностями поведения в популяции, которые существовали до появления в инфекции популяции. Однако, когда инфекция распространяется, люди видят ее последствия и изменяют поведение (снижают число сексуальных партнеров и/или начинают чаще пользоваться презервативами). В результате базовое репродуктивное число меняется и инфекция идет на спад. Однако оно вряд ли снижается настолько, чтобы оказаться ниже единицы, и поэтому распространенность инфекции опять стабилизируется, но на новом, более низком уровне («эндемическая фаза»). Схематически это представлено на рис. 12.

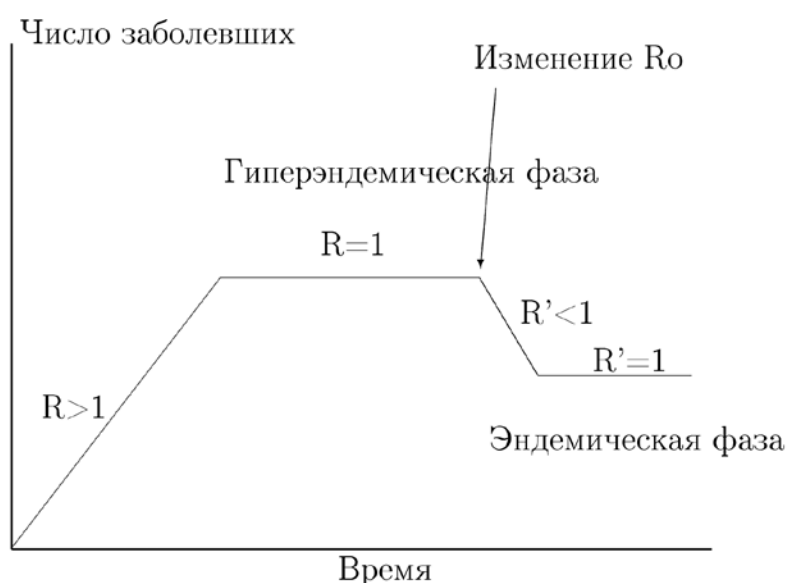


Рис. 12. Схема динамики эпидемического процесса

Чтобы более детально понять, как будет меняться характер эпидемии, предположим, что речь идет о заболевании, передающемся с вероятностью 50% при однократном контакте¹⁷. В популяции в год имеется в среднем 5 контактов (пять сексуальных партнеров), и длительность заболевания без лечения составляет полгода (длительность заразной стадии). В случае лечения этот период сокращается до 1,8 месяца (или 0,15 года — отсрочка связана с несвоевременным выявлением и началом лечения).

¹⁷ Это примерно равно вероятности заражения гонореей.

При попадании инфекции в такую популяцию она будет иметь репродуктивное число, равное $R_0 = 5 * 0,5 * 0,5 = 1,25$. Заболеваемость будет возрастать и достигнет уровня гиперэндемии, равного $I = 1 - 1/1,25 = 0,2$. В такой популяции распространенность инфекции составит 20% от общей численности популяции в каждый момент времени. Однако широкое распространение заболевания приведет к тому, что люди начнут менять свое поведение, например, ранее обращаясь за лечением и/или уменьшая число партнеров.

Предположим, что единственное, что произошло в этой популяции — это увеличение частоты обращаемости за медицинской помощью, в результате которого 20% популяции теперь обращаются вовремя и у них средняя продолжительность заразного периода составляет 0,15 года. Поведение остальной популяции осталось без изменений. Тогда средняя продолжительность заразного периода составит $D = 0,15 * 0,2 + 0,5 * 0,8 = 0,43$. Базовое репродуктивное число снизится до $R_0 = 5 * 0,5 * 0,43 = 1,075$. Число случаев инфицирования начинает снижаться и стабилизируется на уровне 7%. Изменение поведения привело к новому уровню распространенности заболевания.

Знание особенностей распространения инфекции в SIS-моделях позволяет ориентировочно оценивать эффективность профилактических мероприятий и устанавливать требования к степени охвата профилактическими вмешательствами.

Прежде чем двигаться дальше, следует остановиться еще на одной особенности оценки репродуктивного числа. Как известно, базовое репродуктивное число зависит от числа контактов в единицу времени и вероятности заражения за один контакт. В случае инфекций, передающихся при половых контактах, естественной единицей измерения является один половой акт. Однако при проведении социологических опросов участников крайне редко спрашивают о числе половых актов за определенный промежуток времени. Чаще всего спрашивают о количестве сексуальных партнеров (например, за прошедший год). Однако сама по себе эта цифра ничего не говорит, поскольку она не учитывает числа контактов, при которых может произойти заражение. Поэтому аналитику требуется рассчитать вероятность заражения в том случае, если один из сексуальных партнеров окажется инфицированным. Эта величина зависит от числа половых актов за одно партнерство и вероятности заражения за один акт. Например, для ВИЧ-инфекции вероятность заражения при однократном вагинальном половом сношении составляет 1 на 1000 половых актов. Для расчетов вероятности заражения следует знать как минимум частоту половых актов и среднюю длительность партнерства. Так, если средняя частота половых актов составляет 8 в месяц, то за год это составит 96 половых актов. Тогда вероятность заразиться у уязвимого партнера в ВИЧ-дискордантной паре рассчитывается исходя из следующих соображений. Предположим, что есть популяция из 1000 таких дискордантных пар. Если за определенный период времени (год) у них будет только один половой акт, то из 1000 пар заразится только одна. Остальные 999 останутся дискордантными. Однако если будет два половых акта, то во время второго заразится уже 1/1000 оставшихся 999 пар. На следующий этап могут перейти только 99,9% оставшихся дискордантными пар. Соответственно, на каждый последующий этап будут переходить 99,9% оставшихся, а общее число заразившихся будет равно разности числа тех, кто был дискордантным вначале и числа тех, кто останется дискордантным в самом конце. Тогда вероятность заражения за n половых актов составит

$$\beta_x = 1 - (1 - \beta)^n.$$

В описанном выше примере вероятность заражения за один год с дискордантным партнером составит:

$$\beta_x = 1 - (1 - 0,001)^{96} = 0,092 \quad (9,2\%).$$

Для упрощения расчетов следует помнить, что в случае низкой вероятности заражения (когда разность $1 - \beta$ близка к единице), выражение для β_x можно упростить так:

$$\beta_x = 1 - (1 - n * b) = n * b.$$

Пусть продолжительность заразного периода составляет 8 лет, а среднее число партнеров в год составляет 3 человека. Тогда базовое репродуктивное число составит $R_0 = 3 * 8 * 0,092 = 2,2$. Очевидно, что вирус будет распространяться в такой популяции.

Теперь попытаемся выяснить, как широко должно быть распространено использование презервативов, чтобы вирус в такой популяции не распространился. Презервативы не обеспечивают стопроцентной защиты, однако они снижают риск заражения до 0,15 от исходного. Соответственно, при использовании презервативов вероятность заражения (обозначим ее β_1) составит 0,15/1000. Соответственно, точное значение вероятности заражения за один год составит 0,014. При использовании презервативов для каждого полового акта в описанной выше ситуации репродуктивное число составит $R_0 = 3 * 8 * 0,014 = 0,343$.

Обозначим число половых актов, при которых презерватив используется, np , а когда нет — nn . Использование презервативов не влияет на число половых актов, поэтому $nn + np = n$. Если p — пропорция использующих презервативы $p = np/n$, то тогда $np = p * n$, а $nn = (1 - p) * n$.

Пользуясь упрощенной формулой, вероятность заражения за одно партнерство составит:

$$\beta_x = 1 - (1 - np * \beta_1) + 1 - (1 - nn * \beta_0)$$

Или, упрощая,

$$\beta_x = np * \beta_1 + nn * \beta_0 = n * (p * \beta_1 + (1 - p) * \beta_0)$$

Подставляя это выражение в формулу базового репродуктивного числа, получаем:

$$R_0 = c * n * (p * \beta_1 + (1 - p) * \beta_0) * D.$$

Если задачей профилактической программы является предотвращение распространения эпидемии, то с точки зрения влияния на параметры эпидемии ее задачей является уменьшение базового репродуктивного числа (R_0) до уровней, не превышающих единицу. Соответственно, для программ, направленных на стимулирование использования презервативов, описанная выше формула позволяет определить, какая частота их использования будет достаточной для достижения поставленной цели. Принимая $R_0 = 1$ и решая уравнение относительно процента использующих презервативы p , легко обнаружить, что минимальный необходимый уровень использования презервативов определяется следующей формулой:

$$p = \frac{c * n * \beta_0 * D - 1}{c * n * D * (\beta_0 - \beta_1)}$$

Стоит обратить внимание на то, что выражение в числителе представляет собой показатель, демонстрирующий, насколько исходное репродуктивное число было больше единицы (т.е. насколько профилактическая программа должна снизить репродуктивное число для остановки эпидемии). В знаменателе стоит разность репродуктивных чисел до проведения вмешательства

и в той группе, которая выполняет рекомендации (в описываемом примере это лица, которые используют презервативы. Соответственно, эту формулу можно записать в следующей, более простой форме:

$$p = \frac{R_{0b} - 1}{R_{0b} - R_{0i}},$$

где R_{0b} — репродуктивное число без вмешательства, а R_{0i} — репродуктивное число в группе, последовавшей рекомендациям. Описанный выше пример базировался на следующих значениях репродуктивных чисел: $R_{0b} = 2,2$ и $R_{0i} = 0,34$. Соответственно, для того, чтобы остановить развитие эпидемии, необходимо, чтобы презервативами при каждом половом акте пользовались не менее 64% популяции риска:

$$p = \frac{2,2 - 1}{2,2 - 0,34} = 0,64$$

В реальной ситуации люди, утверждающие, что пользуются презервативами, не всегда пользуются ими при каждом половом акте и, соответственно, вероятности заражения будут выше. В принципе, можно попытаться проанализировать выражение для репродуктивного числа для того, чтобы определить, в каком проценте половых актов (в среднем) необходимо добиться использования презервативов для того, чтобы эпидемия перестала прогрессировать. Не приводя детального вывода, полученное выражение для процента индивидуальных контактов составит:

$$p_c = \frac{\frac{1}{N} \ln\left(\frac{c * D}{c * D - 1}\right) + \ln(1 - \beta)}{\ln(1 - \beta) - \ln(1 - k * \beta)},$$

показывая, что процент зависит от общего количества индивидуальных контактов с одним партнером N , количества партнерств c , вероятности заражения при однократном контакте β и того, насколько профилактическое вмешательство снижает вероятность заражения k . Для описанного выше примера для того, чтобы предотвратить дальнейшее распространение эпидемии ВИЧ-инфекции, необходимо, чтобы не менее 59% всех половых актов в данной популяции совершались с использованием презерватива. Располагая данной информацией, можно проводить опросы населения и определять, насколько успешным (или неуспешным) было вмешательство. Если после проведения вмешательства контрольных цифр достичь не удастся, это означает, что основная задача выполнена не была. Более того, даже до начала вмешательства можно оценить возможность успеха и при необходимости использовать многосторонние подходы в том случае, если акцент только на одном типе вмешательства не окажет выраженного воздействия на течение эпидемии.

Репродуктивное число и гетерогенные популяции. Описанные выше подходы предполагали, что изучаемая популяция является достаточно гомогенной. Ряд гетерогенных моделей мы разберем позднее, однако сейчас хотелось бы остановиться на одном из путей преодоления проблемы гетерогенности популяции.

В области моделирования инфекций, передающихся половым путем, включая ВИЧ-инфекцию, основным по важности фактором гетерогенности является количество партнеров. Дело в том, что два других компонента базового репродуктивного числа — вероятность заражения при контакте и длительность заразного периода — определяются в большей степени свойствами возбудителя¹⁸. В то же время количество партнеров определяется самими людьми и, соответственно, по этому

¹⁸ Следует, правда, учитывать, что человек также может влиять на эти параметры, например, путем использования презервативов (влияние на вероятность заражения) или своевременным обращением за лечением (влияние на длительность заразного периода).

параметру люди отличаются друг от друга весьма значительно. Поскольку популяция становится гетерогенной, то использовать описанные в предыдущем разделе формулы уже нельзя. Однако существует два относительно простых подхода к их улучшению (подробнее см. [3]).

Первый подход базируется на замене количества партнеров в описанных выше формулах на т.н. «среднюю эффективную скорость смены партнеров»¹⁹. Эта величина больше средней численности партнеров на величину, пропорциональную дисперсии (т.е. разбросу) численности партнеров в популяции.

$$c_e = \bar{c} + \sigma^2 / \bar{c}$$

Формула говорит о том, что средняя эффективная скорость смены партнеров c_e равна среднегрупповой численности партнеров (\bar{c}), увеличенной на отношение дисперсии численности партнеров (σ^2) к численности партнеров. Дисперсия численности может быть определена путем расчета выборочной дисперсии, по известным из статистики формулам:

$$\sigma^2 = \frac{\sum (c_i - \bar{c})^2}{n - 1},$$

где n — численность выборки, а c_i — количество партнеров у i человека в выборке.

Полученная в результате расчетов величина используется вместо средней численности партнеров в формулах расчетов базового репродуктивного числа и иных показателей, описанных ранее. Приведем в качестве примера оценку средней эффективной скорости смены партнеров в группе 129 женщин, обследованных в рамках популяционного опроса в г. Санкт-Петербурге (И. Полонская, не опубликовано, см табл. 4).

Таблица 4

Численность женщин с разным количеством смен партнеров в течение года

c_i	Число женщин (k)	$(c_i - \bar{c})$	$*(c_i - \bar{c})^2$
0	23	-1,496	51,5
1	69	-0,496	17,0
2	19	0,504	4,8
3	9	1,504	20,4
4	2	2,504	12,5
5	4	3,504	49,1
6	2	4,504	40,6
19	1	17,504	306,4
	1,496		$\Sigma = 502,3$

В этой таблице рассчитано среднее количество смен партнеров в данной группе (оно равно 1,496), а затем проведен расчет дисперсии. Поскольку сумма квадратов отклонений численности партнеров от среднего (последняя колонка табл. 4) равна 502,3, дисперсия будет равна этой величине, деленной на 128 (129 женщин – 1) или 3,92. Соответственно, средняя эффективная скорость

¹⁹ Точнее, скорость приобретения новых партнеров.

смены партнеров будет составлять $1,496 + 3,92/1,496 = 4,12$. Простое сравнение этой величины с исходным значением среднего числа партнеров показывает, что реально репродуктивное число в такой популяции будет почти в три раза выше, чем если бы коррекция на гетерогенность не использовалась.

Следует заметить, что использованные формулы достаточно ясно демонстрируют, что в том случае, если группа гомогенная (различий по числу партнеров нет), формула эффективной средней скорости смены партнеров сводится просто к числу новых партнеров, поскольку в этом случае дисперсия равна нулю.

Другим подходом к оценке влияния гетерогенности групп является расчет средневзвешенного репродуктивного числа. Этот подход позволяет принять во внимание различия не только в среднем количестве партнеров, но и других поведенческих факторах риска, которые могут оказать воздействие на репродуктивное число для данной подгруппы. Единственным сохраняющимся допущением является гомогенность скрещивания, т.е. вероятность контактов между группами является пропорциональной активности этой группы²⁰. Если это так, то для определения среднего репродуктивного числа необходимо подсчитать общее количество новых партнеров, появившихся у данной группы (в табл. 4 это будет произведение c_i на k), а затем оценить, какой процент этих партнеров приходится на ту или иную группу. Затем сумма произведений этого процента на групповое базовое репродуктивное число и даст средневзвешенное репродуктивное число. Для примера опять воспользуемся данными опроса в Санкт-Петербурге, но возьмем суммарную группу мужчин и женщин²¹ (196 человек, табл. 5).

Таблица 5

Численность мужчин и женщин с разным количеством смен партнеров в течение года

c_i	k	$c_i * k$	p	R_0	$p - R_0$
0	30	0	0	0	0
1	102	102	0,305	0,25	0,076
2	30	60	0,180	0,5	0,090
3	15	45	0,135	0,75	0,101
4	5	20	0,060	1,0	0,060
5	6	30	0,090	1,25	0,112
6	3	18	0,054	1,5	0,081
8	1	8	0,024	2,0	0,048
9	1	9	0,027	2,25	0,061
11	1	11	0,033	2,75	0,091
12	1	12	0,036	3,0	0,108
19	1	19	0,057	4,75	0,270
		$\Sigma = 334$			$\Sigma = 1,098$

Из этой таблицы видно, что при выполнении поставленных в основу модели допущений на лиц с большим количеством партнеров приходится непропорционально большое количество контактов. Так, например, лица с количеством партнеров 19 составляют в данной группе только

²⁰ Это допущение, как будет описано ниже, выполняется далеко не всегда и поэтому приходится использовать более сложные модели.

²¹ В реальности надо делать взвешенное репродуктивное число у женщин по численности партнеров-мужчин, а у мужчин — по численности партнеров-женщин. В данном случае используем упрощенные расчеты.

полпроцента, однако на них приходится почти 6% всех контактов. С другой стороны, те лица, которые не меняли в течение года партнеров, не влияют на базовое репродуктивное число в популяции (поскольку они не относятся к популяции риска).

Кроме того, внимательное рассмотрение таблицы указывает еще на один важный фактор. Суммарное базовое репродуктивное число больше единицы (1,098) и, соответственно, в данной популяции изучаемое инфекционное заболевание может распространяться. Как указывалось ранее, уровень распространенности, соответствующий этому базовому репродуктивному числу, составит 8,9%. Предположим теперь, что у нас есть две стратегии профилактики. Одна из них (назовем ее низкорисковая) предлагает скрининг и лечение всем желающим, но не фокусируется на группе риска (лицах с большим количеством партнеров). Поскольку затраты на поиск невысоки, предположим, что она стоит 100 условных единиц, но воспользуются ею только лица из группы низкого риска (с одним и двумя новыми партнерами в год). Таких человек 162, соответственно, стоимость программы составит 16 200 условных единиц. В результате раннего выявления и лечения удастся снизить продолжительность заразного периода в два раза — с 0,5 до 0,25, соответственно, репродуктивное число в группе с одной сменой партнера снизится до 0,125, а в группе с двумя партнерами — до 0,25. Что же произойдет с базовым репродуктивным числом и распространенностью? Легко рассчитать, что базовое репродуктивное число снизится до 1,014, а распространенность — до 1,4%. Распространенность заболевания снизилась, однако она не упала до нуля и для предотвращения нового распространения надо будет ежегодно затрачивать указанные ранее суммы на скрининг и лечение, и при этом около полутора процентов населения будут страдать от этого заболевания.

Теперь обратимся ко второй стратегии профилактики — высокорисковой. Она состоит в выявлении и лечении только лиц высокого риска, например, тех, у кого больше 10 партнеров в год. Предположим, что для них, учитывая сложности с выявлением этих лиц, стоимость одного случая в 10 раз выше. Поскольку всего таких человек трое, общая стоимость программы составит 3000 условных единиц. Индивидуальная эффективность программы такая же, как и в случае низкорисковой профилактики, т.е. период заразности сокращается в два раза. Легко рассчитать, что тогда репродуктивное число в этих группах составит 1,375, 1,5 и 2,375, соответственно. Суммарное же репродуктивное число упадет до 0,86. Это означает, что в такой популяции распространенность снизится до нуля.

Итак, низкорисковая стратегия профилактики позволяет снизить распространенность на с 8,9% до 1,4% за 16 200 условных единиц, а высокорисковая — с 8,9% до 0 за 3000 единиц. Это означает, что один предотвращенный случай в группе низкорисковой профилактики стоит 2160 условных единиц, а в группе высокорисковой профилактики — 337 условных единиц. Более того, с каждым годом этот эффект будет нарастать.

Этот пример хорошо иллюстрирует, почему высокорисковая стратегия профилактики является более стоимостно-эффективной: группы высокого риска являются менее многочисленными, но на них приходится большее количество опасных контактов. Поэтому с точки зрения обеспечения общественного здоровья, здоровья всей популяции, необходимо обеспечивать здоровье групп риска. Если средства будут тратиться на группы риска, то и менее рискованные группы будут более здоровыми.

В целом же приведенные выше данные показывают, что в рамках SIS-моделей можно проводить достаточно полноценный анализ распространения заразных заболеваний, однако эти модели не всегда полностью описывают эпидемический процесс, что требует использования других моделей, описанных ниже. Следует, однако, указать, что в случаях открытых популяций многие другие модели превращаются в SIS-модели, и поэтому обсужденные выше подходы к оценке

базового репродуктивного числа становятся важными для подавляющего числа инфекционных заболеваний человека.

3.1.3. SIR-модели

Во многих случаях после того, как инфекционное заболевание завершается, человек не может больше им заразиться — в большинстве случаев это может быть связано с тем, что пациент приобретает иммунитет, но в ряде случаев — и поскольку пациент умирает. В любом случае он более не заразен, но и не возвращается в пул уязвимых. Он удаляется, по-английски *Removed*, откуда следует название моделей — *SIR* (*Susceptible-Infected-Removed*). Оригинально подобные модели расшифровывались как уязвимые-инфицированные-резистентные (*Resistant*), но для того, чтобы можно было описывать модели со смертностью пациентов, расшифровка последнего *R* в названии моделей изменилась на удаленные (мрачнее, но правильнее было бы называть эти модели *SID* — *Susceptible-Infected-Dead* — Уязвимый-Инфицированный-Мертвый). Вообще-то те *SIR*-модели, которые предполагают выздоровление и формирование иммунитета у пациента, отличаются от *SID*-моделей в том, что в первых численность популяции не меняется, и поэтому вероятность заражения с течением времени снижается (по мере того, как все больше пациентов переходят в стадию выздоровевших с иммунитетом). В случае же *SID*-моделей снижается численность популяции, при этом вероятность заразного контакта снижается значительно медленнее (поскольку «разбавляющего» действия лиц с иммунитетом нет). В *SIR*-моделях возможна ситуация, когда уязвимый пациент, находящийся в популяции, не заражается, несмотря на вспышку высокозаразного заболевания, поскольку его контакты представлены в основном переболевшими или иммунными сородичами. Подобный феномен называется стадным иммунитетом (*herd immunity*) и лежит в основе расчетов процента покрытия вакцинацией. В *SID*-моделях прослойки иммунных нет и, соответственно, стадный иммунитет уязвимых не защищает. Не случайно поэтому *SID*-модели в открытой популяции часто лучше аппроксимируются *SIS*, а не *SIR*-моделями.

В классической *SIR*-модели пациент изначально находится в уязвимом состоянии (он здоров, и у него нет иммунитета). Он контактирует с заразным пациентом, который может ему передать инфекцию с вероятностью β . Каждый заразный пациент совершает с контактов в единицу времени. Соответственно, процент контактов, во время которых происходит заражение, при контакте с полностью уязвимой популяцией составляет $c * \beta$. Однако в популяции присутствуют как уязвимые пациенты, так и заразные и выздоровевшие (и имеющие иммунитет). Соответственно, количество контактов с уязвимыми пациентами пропорционально количеству уязвимых в популяции в целом $\frac{S}{S+I+R}$, где *R* — количество иммунных в популяции).

Следует обратить внимание на то, что в *SID*-моделях $R = 0$. Таким образом скорость изменения численности уязвимых определяется уравнением:

$$S_t = -\beta * c \frac{S}{S+I+R} * I$$

Знак минуса показывает, что количество уязвимых убывает. Аналогичным образом прирост числа инфицированных определяется той же формулой (без отрицательного знака), однако, в отличие от уязвимых, на численность инфицированных влияет и второй процесс — убыль в результате выздоровления (или смерти в *SID*-моделях). Соответственно, для количества инфицированных формула приобретает вид:

$$I_t = \beta * c \frac{S}{S+I+R} * I - \frac{I}{D}$$

До этого места модель практически аналогична обычной SIS-модели. Разница проявляется на следующем этапе, когда появляется группа выздоровевших и иммунных (или умерших в SID-моделях). Обозначим эту группу R (Removed или Resistant). Очевидно, что численность этой группы меняется со скоростью, равной скорости окончания заболевания (выздоровления или смерти). Соответственно, эта скорость равна

$$R_i = \frac{I}{D}$$

Установив эти взаимоотношения между показателями, можно представить модель в графическом виде с помощью системы Vensim. Для этого возьмем описанную выше (рис. 2) SIS-модель. В ней необходимо устранить обратный поток выздоравливающих от инфицированных к уязвимым и «перенаправить» его к новому контейнеру — выздоровевшим. Соответственно, численность выздоравливающих продолжает зависеть от численности инфицированных и длительности заболевания. Кроме того, очевидно, что вероятность встречи с уязвимым зависит не только от численности уязвимых и инфицированных, а от общего размера популяции (суммы уязвимых, инфицированных и выздоровевших). Поэтому необходимо добавить стрелку от выздоровевших к потоку «заражения». Граф модели выглядит так:

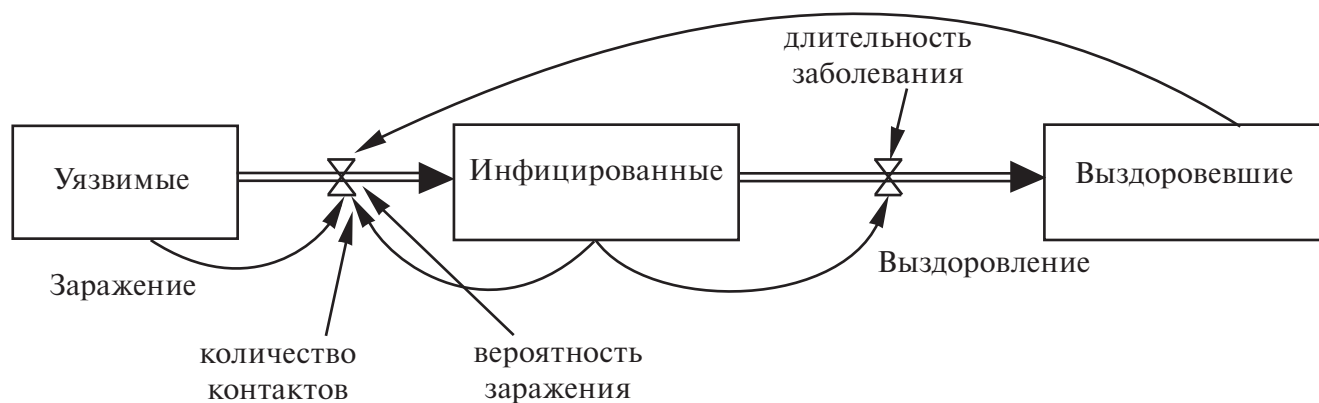


Рис. 13. Граф SIR-модели

Эту модель можно проанализировать точно так же, как ранее это делалось для SIS-моделей. Представим результаты для численности инфицированных пациентов (рис. 14).

Видно, что эпидемия имеет ограниченную продолжительность, пик достигается на 30-й день, а затем численность инфицированных снижается, пока вспышка не прекращается к 85-му дню с момента ее начала. Если при этом рассмотреть изменения численности не болевших, то выяснится, что к 50-му дню таковых не остается, и поэтому вспышка идет на убыль. Кажется, что она просто «выгорает», больше не остается людей, которые могли бы заболеть. Однако, если рассматривать динамику внимательно, то становится понятно, что вспышка начинает угасать еще до того, как все люди в популяции переболеют этой инфекцией. Действительно, к пятидесятому дню, когда в популяции не остается уязвимых лиц, количество инфицированных также уже незначительное. Соответственно, можно предположить, что если бы заболевание было менее заразным (в нашем примере базовое репродуктивное число $R_0 = 5 * 0,1 * 10 = 5$), то могли бы остаться лица, которые не заразились.

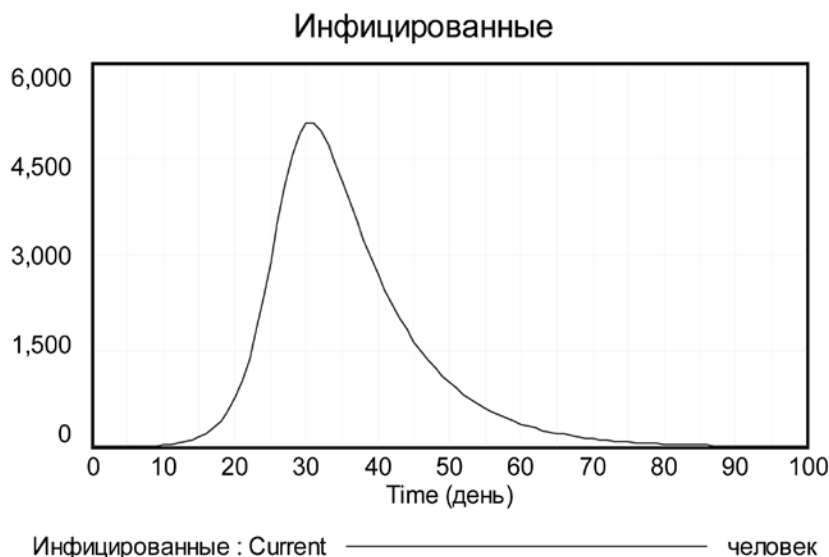


Рис. 14. Результаты анализа SIR-модели с параметрами, аналогичными описанными выше для SIS-модели (см. рис. 10)

Соответственно, можно попробовать немного изменить условия моделирования и снизить, например, заразность заболевания. Если вероятность заражения при однократном контакте снизится в три раза, базовое репродуктивное число все еще будет выше единицы (составляя примерно 1,7) и, соответственно, можно ожидать возникновения вспышки. Действительно, созданная модель предсказывает возникновение вспышки, правда, развивающейся значительно медленнее (пик количества инфицированных приходится на 125–130 день со дня начала эпидемии). Вместе с тем анализ численности уязвимых показывает, что почти 35% уязвимой популяции так и не заразится, несмотря на то, что они также находились в группе риска и, согласно определению модели, ничем не отличались от тех, кто заразился. Они оказались защищены «стадным» иммунитетом²². Продемонстрируем этот феномен при помощи SIR-модели в SAS.

Для этого возьмем модель SIS, описанную выше, и немного модифицируем ее (создание файла с параметрами длительности моделирования приводить не будем, поскольку там никаких изменений нет (кроме удлинения времени моделирования), так же, как и не меняются команды создания графика).

1. PROC MODEL DATA=t;
2. DEPENDENT Inf 1 Sus 10000 Res 0;
3. PARMS c 5 beta 0.033 D 10;
4. Infected=c*beta*Inf*Sus/(Inf+Sus+Res);
5. Cured=Inf/D;
6. DERT.Inf=Infected–Cured;
7. DERT.Sus=–Infected;
8. DERT.Res=Cured;
9. SOLVE Inf Sus Res/ OUT=determ;
10. RUN; QUIT;

²² Понятно, что в SID-моделях группа удаленных не «разбавляет» популяцию, поэтому в такой ситуации вспышка угасает, только когда в нее оказываются втянутыми все уязвимые.

Изменения в программе начинаются с появления дополнительной зависимой переменной (аналогично тому, как в модели Vensim появился новый контейнер). Мы назвали эту переменную Res (от английского ‘Resistant’ — устойчивые). В нулевой момент времени в популяции устойчивых нет²³. Появилось и новое уравнение, которое показывает, что резистентными становятся те, кто выздоровел (Cured). Фактически вся модель сводится лишь к трем простым выражениям, показывающим, что количество инфицированных увеличивается за счет заразившихся и уменьшается в результате выздоровления. Количество уязвимых уменьшается в результате заражения, а резистентные — это те, кто выздоровел.

Полученный в результате анализа этой модели график приведен на рис. 15.

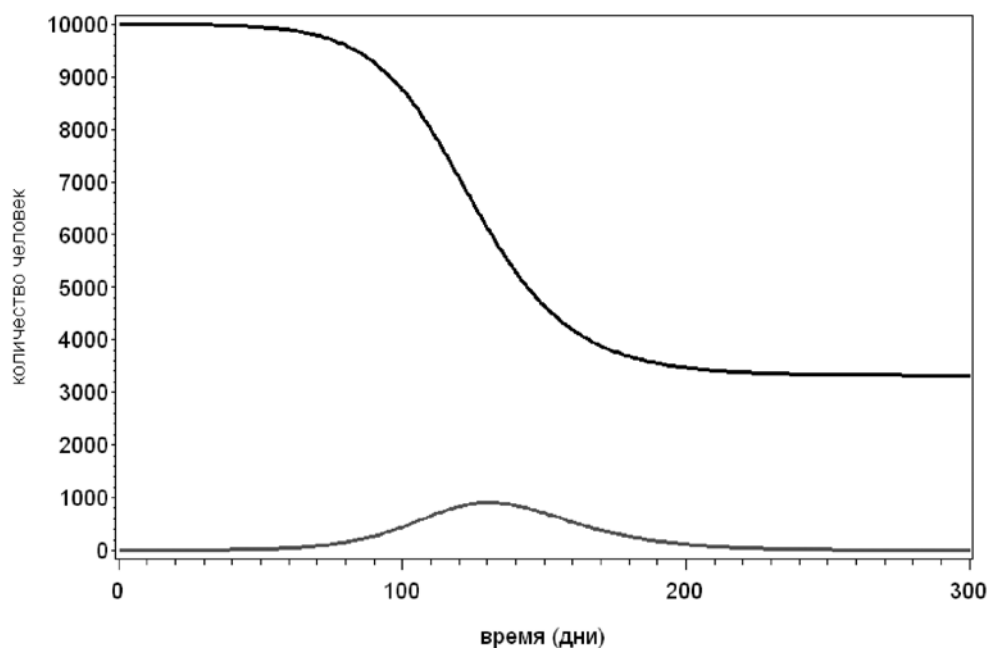


Рис. 15. Результаты анализа SIR-модели в SAS

График, построенный по результатам модели, оказывается значительно более информативным, чем отдельные графики количества уязвимых и инфицированных, которые генерирует учебная версия Vensim. Четко видно, что большая часть населения не вовлекается в эпидемический процесс. В любой момент времени распространенность инфекции (количество инфицированных по отношению к общей численности популяции) остается достаточно низким, не повышаясь выше 10%, хотя общее число переболевших и составляет более 60% популяции. Однако наиболее интересным является описанный выше феномен достаточно большого количества уязвимых пациентов, остающихся в популяции к моменту окончания вспышки.

Поскольку этот феномен существует, чрезвычайно важно знать, какой уровень «стадного» иммунитета может в принципе подавить возникновение вспышки. В принципе, этот вопрос можно решить, используя «грубую силу» компьютерного моделирования, меняя входные параметры и отмечая тот момент, когда модель перестанет предсказывать возможность развития вспышки²⁴. Однако попробуем решить ее аналитически.

²³ Изменение исходной численности устойчивых полезно при изучении вопроса о влиянии иммунизации на течение эпидемического процесса.

²⁴ Этот подход будет являться разумным, если модель достаточно сложная.

Выше уже отмечалось, что количество зараженных в единицу времени определяется соотношением:

$$I_t = \beta * c * I \frac{S}{S + I + R}$$

В то же время скорость выздоровления определяется как

$$R_t = \frac{I}{D}$$

Для того, чтобы вспышка пошла на убыль необходимо, чтобы количество выздоравливающих превышало количество заражающихся, т.е.

$$\frac{I}{D} > \beta * c * I \frac{S}{S + I + R}$$

или

$$\beta * c * D \frac{S}{S + I + R} < 1$$

Произведение $\beta * c * D$ — это не что иное, как уже неоднократно упоминавшееся выше базовое репродуктивное число, соответственно, если обозначить процент уязвимых в популяции буквой p , то формула становится очень простой:

$$R_0 * p < 1$$

и из нее следует вывод, что процент уязвимых в популяции, необходимый для того, чтобы вспышка пошла на убыль (или не развилась), должен быть меньше, чем величина, обратная базовому репродуктивному числу для данного заболевания. Соответственно, для заболевания с репродуктивным числом 5 после падения численности уязвимых ниже 20% вспышка пойдет на убыль, а для заболевания с репродуктивным числом 1,7 для этого потребуется, чтобы численность уязвимых снизилась ниже 58%. Действительно, если, например в описанной выше модели в SAS указать, что из 10 000 человек 4200 являются резистентными, а 5800 — уязвимыми, модель перестает предсказывать развитие вспышки.

Это простое соотношение позволяет легко оценивать потребность в покрытии вакцинацией для предотвращения вспышек. Поскольку до момента заноса инфекции в популяции, где возможна вакцинопрофилактика, существуют только уязвимые (невакцинированные) и резистентные (вакцинированные), то, соответственно, количество вакцинированных, которые должны присутствовать в популяции, должно быть больше разности единицы и величины, обратной базовому репродуктивному числу.

$$Vac = 1 - \frac{1}{R_0}$$

Используя это соотношение, видно, что для инфекций с базовым репродуктивным числом, равным 5, необходим 80%-ный охват вакцинацией, а для инфекций, например, с базовым репродуктивным числом 15 вакцинировано должно быть 93% популяции.

На самом деле вакцинация снижает количество заразных контактов, она никак не влияет на вероятность заражения для уязвимого (по определению) и не влияет на скорость выздоровления (опять же по определению). Поэтому единственный компонент в формуле базового репродуктивного числа, который меняется — это количество контактов s . Именно на это число, кстати, и направлены многие противоэпидемические мероприятия. Действительно, например, отмена школьных занятий во время эпидемии гриппа не устраняет полностью контакты между детьми, однако она достаточно резко снижает их, влияя таким образом на репродуктивное число.

Описанная выше модель предполагала, что, раз попав в группу резистентных, люди навсегда остаются в ней. Хотя это допущение и может быть справедливо для ряда инфекций, оно достаточно плохо описывает реальную ситуацию в популяции. Дело в том, что количество уязвимых в популяции постоянно пополняется, и происходит это за счет миграции населения и рождений.

Поэтому простейшие SIR-модели хороши только для описания отдельных вспышек заболеваний, продолжительность которых относительно продолжительности жизни человека относительно невысока. Однако, если мы хотим проанализировать длительную динамику, то не учитывать рождаемость (и смертность среди резистентных) нельзя. Такие популяции, в которых население может мигрировать и из которых оно может выбывать, называются открытыми.

3.1.4. Модели открытых популяций

В принципе модели для открытых популяций должны включать как минимум два параметра — рождаемость и смертность. Однако очень часто для простоты считают, что рождаемость и смертность равны друг другу и исследователь имеет дело со стационарной популяцией, т.е. популяцией, численность которой остается постоянной во времени.

Проще всего стационарная популяция моделируется путем приравнивания рождаемости смертности и, хотя подобный подход является явным упрощением, он позволяет получить ряд достаточно интересных результатов. Для того, чтобы построить SIR-модель для открытой стационарной популяции, вначале модифицируем граф простейшей SIR-модели так, как показано на рис. 16.

Как видно на рис. 16, в модель был добавлен один контейнер (умершие) и два соответствующих ему потока из группы выздоровевших и уязвимых (в принципе можно было бы добавить еще один поток — от инфицированных, однако в том случае, если продолжительность заболевания относительно невелика, смертностью инфицированных можно пренебречь²⁵. Смертность как выздоровевших, так и уязвимых определяется продолжительностью жизни в популяции (которая опять-таки упрощенно моделируется как имеющая экспоненциальное распределение). Соответственно, два потока (умершие уязвимые и умершие резистентные) зависят от продолжительности жизни и численности соответствующих исходных групп. Количество умерших в единицу времени определяется как отношение количества лиц в группе к продолжительности жизни.

Для того, чтобы сделать популяцию стационарной, рождаемость была приравнена к смертности, поэтому рождаемость соединена стрелочками с обоими показателями смертности. Следует обратить внимание на то, что поток рождений идет «из ниоткуда». Это специальный символ, показывающий, что исходное значение нам либо неизвестно, либо оно нас не интересует. В прин-

²⁵ Конечно, речь идет о заболеваниях с низкой летальностью, если моделируемое заболевание имеет высокую летальность, без дополнительного потока не обойтись.

ципе, можно было бы не делать и контейнера «умершие», а сделать потоки смертности также идущими «в никуда». Это можно делать, когда основной интерес направлен на изучение динамики численности инфицированных или переболевших лиц. Однако в большом числе приложений аналитика интересует, сколько человек умирает в данной популяции, и для того, чтобы проиллюстрировать, как это делается, в данном графе был оставлен контейнер.

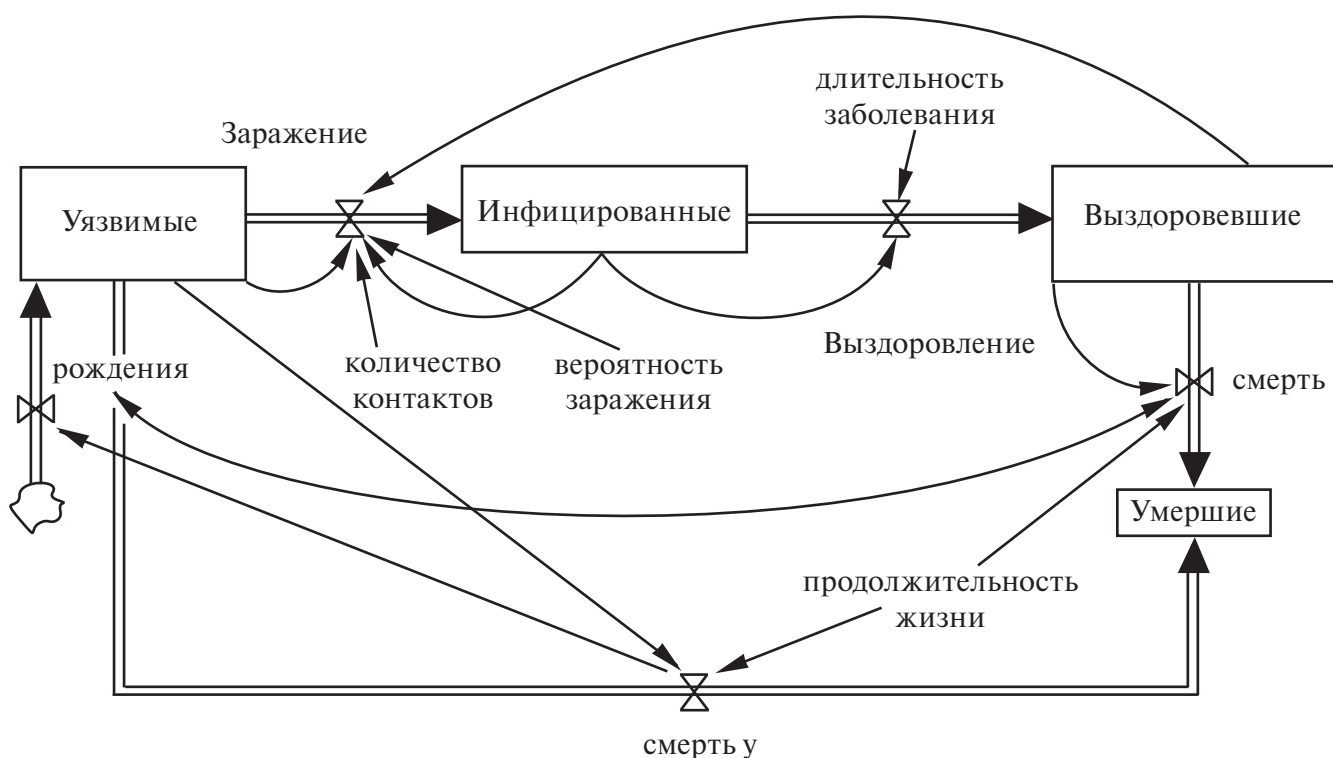


Рис. 16. Граф SIR-модели в открытой стационарной популяции.

Однако при переводе графа в систему SAS мы не будем учитывать количество умерших и, соответственно, у нас по-прежнему останется три контейнера, которые, напомним, в системе SAS обозначаются приставкой DERT. Для перевода модели на язык SAS нам вначале потребуется описать все потоки. Их пять:

1. Заражение
2. Выздоровление
3. Рождение
4. Смерть уязвимых
5. Смерть переболевших

Первые два потока были описаны еще при создании SIS-модели. Поток рождений по определению стационарной популяции равен количеству умерших, т.е. сумме умерших уязвимых и умерших переболевших. Однако количество уязвимых определяется суммой родившихся, инфицированных и умерших, т.е. при расчете количества уязвимых мы отнимаем смертность в этой группе (умершие) и тут же прибавляем ее (в виде родившихся). Понятно, что это можно упростить и анализировать только смертность переболевших. Тогда рождаемость будет равна смертности переболевших и модель упрощается до трех потоков. С этим упрощением программа, описывающая SIR-модель в открытой стационарной популяции, будет выглядеть так:


```

1. PROC MODEL DATA=t;
2.     DEPENDENT Inf 1 Sus 100000 Res 0;
3.     PARMs c 5 beta 0.02 D 50 LE 18250;
4.     Infected=c*beta*Inf*Sus/(Inf+Sus+Res);
5.     Cured=Inf/D;
6.     Dead_r=Res/LE;
7.     DERT.Sus=-Infected+Dead_r;
8.     IF Inf>=0 THEN DERT.Inf=Infected-Cured;
9.     IF Res>=0 THEN DERT.Res=Cured-Dead_r;
10.    SOLVE Inf Sus Res/ OUT=determ;
11.    RUN; QUIT;

```

В приведенной выше программе вначале описаны три потока (инфицированные, выздоровевшие и умершие), а затем три контейнера. Следует обратить внимание на то, что перед контейнерами инфицированных и выздоровевших внесена проверка на наличие в этой группе хотя бы нескольких человек. Система не знает, что речь идет о людях, и поэтому может начать давать отрицательные значения для количества инфицированных и выздоровевших. Для предотвращения этого и используется конструкция IF ... THEN.

Для моделирования в этом примере выбрана продолжительность жизни, равная 50 годам, продолжительность заболевания составляет 50 дней, и репродуктивное число 5. Результаты оценки этой модели приведены на рис. 17, причем для удобства восприятия дни переведены на рисунке в года.

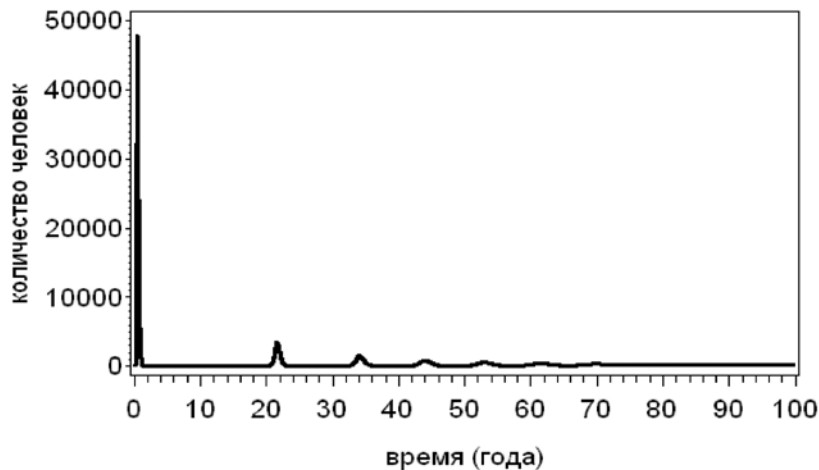


Рис. 17. Количество инфицированных в стационарной открытой популяции, SIR-модель

Как видно на рис. 17, вначале в популяции наблюдается вспышка, причем количество заразившихся является наибольшим из всех последующих вспышек. Далее достаточно длительный период проходит без эпидемий (хотя низкий уровень персистенции возбудителя остается). Затем, когда в популяции формируется достаточно большая прослойка уязвимых (за счет рождения), возникает новая вспышка. До этой вспышки от первой проходит почти 20 лет, и она по своим размерам значительно меньше, чем первая. Далее вспышки повторяются, однако каждый раз они меньше, чем предыдущие. Чрезвычайно важно, что эта модель не делала предположений о селективной смертности переболевших, изменении вирулентности возбудителя или еще каких-то других изменениях. Эта картина повторных вспышек с уменьшающимся процентом

охвата является естественным следствием использования SIR-модели в открытой стационарной популяции. Спустя некоторый (достаточно длительный) период времени вспышки прекращаются и инфекция стабилизируется на низком уровне персистенции в популяции. Таким образом после длительного периода осцилляций SIR-модель в открытой стационарной популяции превращается в SIS-модель.

Однако можно возразить, что данная модель совсем неадекватно описывает реальную ситуацию, поскольку в реальности пополнение популяции идет за счет детей, а они чаще взаимодействуют с другими детьми. Соответственно, ключевое допущение всех описанных выше моделей — случайность контактов между всеми членами популяции — не является правдоподобным. Отклонения от допущения случайности контактов обозначаются терминами ассортативность и дисассортативность. Под ассортативными контактами понимают такие, когда члены одной популяционной группы или имеющие определенные характеристики чаще взаимодействуют друг с другом, чем с членами других популяционных групп. Если, например, рассматривать детей как определенную популяционную группу, то очевидно, что для детей школьного возраста контакты являются в большей степени ассортативными, поскольку они встречаются ежедневно с большим количеством детей, чем взрослых. Однако этот тип контактов не будет справедливым для маленьких детей, которые чаще общаются со взрослыми, чем с другими детьми (очевидным исключением являются дети, посещающие ясли и другие дошкольные детские учреждения). Если количество контактов с членами своей популяционной группы меньше, чем с членами другой, то такой тип контактов называется дисассортативным. Предельным случаем ассортативности является полная независимость одной популяционной группы от другой. Соответственно, чем выше ассортативность группы и чем выше количество контактов внутри нее, тем быстрее будет распространяться заразное начало в ней с относительно меньшим влиянием на популяцию в целом. При наблюдении за популяцией в целом отмечается весьма быстрый рост количества случаев в популяции по мере того, как возбудитель распространяется по одной из популяционных групп. Однако вскоре количество лиц, которые могут быть заражены в этой группе, снижается до того уровня, что более распространение эпидемического процесса становится невозможным (обычное протекание эпидемии по SIR-модели), и количество инфицированных начинает быстро снижаться. Если при этом в популяции присутствует достаточное количество уязвимых, то, в конце концов, за счет контактов между группами эпидемия перебрасывается на более многочисленную популяционную группу и начинается новый рост количества инфицированных, однако на сей раз речь может идти о значительно большей популяционной группе и, соответственно, значительно большей вспышке.

Выделение групп с высокой ассортативностью контактов и большей склонностью к распространению инфекционного возбудителя играет большую роль в эпидемиологии ВИЧ-инфекции. Отдельные популяционные группы, такие, как потребители инъекционных наркотиков, имеют большое количество контактов внутри группы (совместное использование шприцев, игл и другого инъекционного инструментария), поэтому вероятность распространения ВИЧ-инфекции в их среде значительно выше. Поскольку вероятность распространения ВИЧ-инфекции в их среде выше, то, если вирус попадает в популяцию, он начинает в первую очередь распространяться именно в этой группе (а также в группе с большим количеством сексуальных контактов — группе лиц, предлагающих услуги коммерческого секса²⁶). Поэтому эти группы стали объектом т.н. дозорного эпиднадзора (sentinel surveillance). Термин «дозорный эпиднадзор» пришел в эпидемиологию из горного дела, где в качестве «дозорных» в шахтах использовали канареек, которые, будучи более чувствительными к метану, погибали, если в шахте скапливалось его достаточное количество, и люди понимали, что из шахты необходимо эвакуироваться²⁷.

²⁶ Стоит, однако, обратить внимание, что в группе лиц, предлагающих коммерческий секс, инфекция не будет распространяться столь же быстро, что и среди потребителей инъекционных наркотиков, поскольку для первых характерен дисассортативный тип контактов, а для вторых — именно ассортативный.

²⁷ Надо отметить, что сейчас термин «дозорный эпиднадзор» включает в себя не только отслеживание ситуации в группах риска, но и исследования поведенческих факторов риска в попытке предсказать возможный сценарий развития событий.

Аналогичным образом дозорный эпиднадзор полагается на анализ ситуации в группах риска для того, чтобы определить, насколько высока вероятность перехода инфекции из групп риска (групп с высоким риском распространения инфекции и высокой ассортативностью) в общую популяцию. Очевидно, что просто наличие группы риска недостаточно для появления инфекции в общей популяции. Например, для ВИЧ-инфекции тот путь передачи, который является основным для потребителей инъекционных наркотиков, просто отсутствует в популяции. Соответственно, для общей популяции количество контактов с таким путем передачи равно нулю и, соответственно, вирус распространяться не может. Однако в общей популяции вирус может распространяться половым путем, соответственно, необходимы люди, которые бы являлись одновременно членами двух популяционных подгрупп (группы, которая живет половой жизнью (общей популяции), и употребляли бы наркотики). Поскольку, как указывалось выше, ВИЧ является относительно мало заразным заболеванием, эти люди должны относиться к подгруппе с высоким уровнем сексуальной активности. Такие группы называют связующими группами (bridging population, популяции «моста») и к ним в первую очередь относят в эпидемиологии ВИЧ-инфекции работниц коммерческого секса (РКС), являющихся потребителями инъекционных наркотиков. Поскольку РКС являются группой с дизассортативными контактами, они являются хорошим проводником для ВИЧ-инфекции.

Из приведенных выше рассуждений следует, что реалистичная модель распространения ВИЧ-инфекции в популяции должна включать как минимум три популяционных группы (потребители инъекционных наркотиков, работницы коммерческого секса и общая популяция). На самом деле и общая популяция не является однородной по количеству сексуальных контактов, поэтому модель должна учитывать еще и эти факты. Поскольку такая модель будет являться достаточно сложной, вернемся назад к примеру детской инфекции, передающейся воздушно-капельным путем. Для создания модели этого заболевания мы можем выделить две основные популяционные группы — дети и взрослые. При этом вероятность контактов взрослых со взрослыми выше, чем с детьми, а детей с детьми выше, чем со взрослыми (группы имеют ассортативные контакты).

Соответственно, чем же будет определяться распространение инфекции в этой популяции? Можно начать с предположения (неправдоподобного), что между популяционными группами контакты вообще отсутствуют. Тогда модель будет состоять из двух SIR-субмоделей, абсолютно аналогичных разобранным ранее. Естественно, в каждой группе будет своя частота контактов и может быть своя продолжительность заболевания и вероятности заражения при однократном контакте (однако в дальнейшем будем предполагать, что в этом дети и взрослые не различаются, хотя количество контактов между детьми выше). Группы связаны друг с другом тем, что дети взрослеют и становятся членами группы взрослых, и начинают вести себя как взрослые (в смысле количества контактов и их распределения между группами). Так же, как это делалось ранее, будем считать скорость взросления равной обратной продолжительности периода «детства». Как уже было описано выше, взрослые вымирают со скоростью, обратно пропорциональной продолжительности жизни. Для сохранения допущения о стационарности численности популяции предполагается, что рождается столько же людей, сколько умирает взрослых.

Для начала модифицируем немного код, который создает файл, определяющий продолжительность периода моделирования:

1. DATA t;
2. years=365*10;
3. DO time=1 TO years;
4. OUTPUT;
5. END;

6. DROP years;
7. RUN;

Вторая строка автоматически пересчитывает количество лет в количество дней, и этот показатель используется для заполнения файла. Поскольку для моделирования вспомогательная переменная years не нужна, она из файла удаляется (команда DROP).

Далее можно переходить к написанию собственно кода для моделирования. Выглядит он так:

```

1. PROC MODEL DATA=t;
2.     DEPENDENT Inf 0 Sus 90000 Res 0 Inf_c 1 Sus_c 10000 Res_c 0;
3.     PARMS c 6 beta 0.005 D 20 LE 20075 c_ca 1 c_c 30 growing 6205;
4.     * взрослые;
5.     Infected=Sus*beta*(c*Inf/(Inf+Sus+Res)+
6.     c_ca*Inf_c/(Inf_c+Sus_c+Res_c));
7.     Cured=Inf/D;
8.     Dead_r=Res/LE;
9.     Dead_s=Sus/LE;
10.    * дети;
11.    Infected_c=Sus_c*beta*(c_ca*Inf/(Inf+Sus+Res)+
12.    c_c*Inf_c/(Inf_c+Sus_c+Res_c));
13.    Cured_c=Inf_c/D;
14.    Adult_s_c=Sus_c/growing;
15.    Adult_r_c=Res_c/growing;
16.    * взрослые;
17.    IF Sus>=0 THEN DERT.Sus=-Infected+Adult_s_c-Dead_s;
18.    IF Inf>=0 THEN DERT.Inf=Infected-Cured;
19.    IF Res>=0 THEN DERT.Res=Cured-Dead_r+Adult_r_c;
20.    * дети;
21.    IF Sus_c>=0 THEN
22.    DERT.Sus_c=-Infected_c-Adult_s_c+Dead_r+Dead_s;
23.    IF Inf_c>=0 THEN DERT.Inf_c=Infected_c-Cured_c;
24.    IF Res_c>=0 THEN DERT.Res_c=Cured_c-Adult_r_c;
25.    SOLVE Inf Sus Res Inf_c Sus_c Res_c/ OUT=determ;
26. RUN; QUIT;

```

Модель как бы состоит из двух SIR-моделей, разобранных ранее. Командой DEPENDENT устанавливается численность групп уязвимых, инфицированных и резистентных в начальный момент времени, как для взрослых, так и для детей (для детей добавлен суффикс _c от английского 'child' — ребенок). Команда PARMS описывает переменные, использованные в моделировании. Вероятность заражения (beta), продолжительность заболевания (D), продолжительность жизни взрослого человека (LE, в днях), продолжительность периода детства (growing, в днях), а также количество контактов между взрослыми (c), между детьми и взрослыми (c_ca) и между детьми (c_c).

Потоки выздоравливающих определяются точно так же, как это делалось ранее. Потоки умерших взрослых снижают численность взрослого населения, а потоки взросления (Adult) — увеличивают. Наибольшие изменения отмечаются в формулах для потоков инфицированных, где вероятность инфицирования зависит не только от числа контактов с больными членами своей группы (детей с детьми и взрослых со взрослыми), но и от количества контактов с другой группой. При этом общее количество контактов нормируется на процент инфицированных из них путем умножения числа контактов на численность инфицированных в другой группе и делением на общую численность этой другой группы.

Для упрощения модели предполагается, что заболевание имеет нулевую смертность и настолько коротко, что «взрослением» за этот период можно пренебречь.

В команде SOLVE были добавлены имена трех новых показателей (контейнеров) для детей. Результаты моделирования для короткого периода времени (один год) в полностью уязвимой популяции показаны на рис. 18.

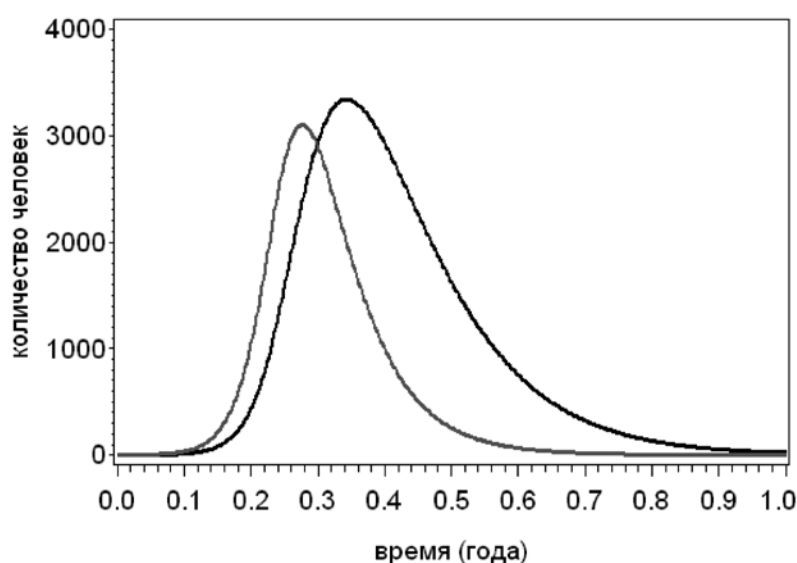


Рис. 18. SIR-модель с двумя популяционными группами и высокой степенью ассортативности контактов

Как видно на рис. 18, вспышка начинается с группы детей (первая линия) и лишь позднее поражает взрослых. Общее количество заболевших детей и взрослых на пике эпидемии примерно равно, хотя численность взрослого населения значительно большая. Ввиду большого количества контактов между детьми заболевание распространяется значительно эффективнее в детской популяции, нежели во взрослой. Интересно отметить, что в представленной выше модели вспышка начинается с заноса одного случая инфекции в детскую популяцию. Однако если изменить модель и указать, что занос произошел во взрослую популяцию, тем не менее, инфекция быстро перебросится на детское население и пик количества инфицированных среди детей будет достигнут все равно ранее, чем среди взрослых. Кроме того, следует обратить внимание, что среди взрослых больше выражен «хвост» эпидемической кривой — она медленнее идет на спад. Это связано с тем, что к моменту достижения пика количества инфицированных численность уязвимых взрослых продолжает оставаться достаточно большой, и поэтому и процесс передачи продолжается дольше, чем среди детей.

Если увеличить продолжительность времени моделирования, то картина будет аналогичной показанной на рис. 17 с рядом постепенно «затухающих» волн, которые иницируются в детской популяции, а затем перебрасываются на взрослую.

Таким образом, меняя переменные, описывающие заразность заболевания, моделируя влияние вмешательств на продолжительность заразного периода и количество контактов, можно прогнозировать течение эпидемического процесса и описывать возможные эпидемические кривые в открытой популяции.

3.1.5. Как улучшить модели?

Выше были описаны основные типы детерминистских моделей, используемых в математической эпидемиологии — SIS-модели, SIR-модели и модели для открытых популяций. Представлены были лишь простейшие модели, включавшие одну-две популяционные группы. Вместе с тем понятно, что, используя все те же блоки, из которых были построены эти модели, можно создавать более сложные и более реалистичные модели. Так, например, можно анализируя SIS-модель инфекции, передаваемой половым путем, сделать четыре группы (по две группы мужчин и женщин — с большим и малым количеством контактов между ними). Если затем оценить степень ассортативности или дисассортативности контактов между группами, то модель будет достаточно хорошо отражать реальность.

Аналогичным образом SIR-модель открытой популяции можно было бы улучшить, поделив взрослую популяцию на подгруппы в зависимости от частоты контактов с другими подгруппами. В конце концов, подобный анализ может привести к достаточно детальному представлению взаимодействий между людьми, как это было сделано, например, в модели EpiSIMS [2].

Еще одной проблемой, которую можно решить путем относительно небольших изменений в приведенных программах, является неопределенность используемых в моделях параметров. Решением этой проблемы является создание разных моделей с разными стартовыми параметрами, и затем изучение уже не одной эпидемической кривой, а их группы с определением наиболее и наименее вероятных сценариев развития ситуации.

Достигается это, например в SAS, путем помещения программного кода модели в макропоследовательность, которая каждый раз перед запуском меняет переменные в модели (например, количество контактов или вероятность заражения). Полученные данные затем анализируются как любые суммарные данные.

Особенно полезным данный подход может оказаться тогда, когда необходимо анализировать одновременно влияние большого количества параметров, каждый из которых имеет свои пределы неопределенности. Тогда можно случайным образом выбирать из диапазона возможных значений одно и (повторив процедуру для всех параметров) использовать сгенерированный таким образом уникальный набор параметров при запуске модели. Количество запусков модели должно быть достаточно большим, чтобы проанализировать большую часть возможных сочетаний параметров. Полученный набор данных может затем анализироваться так, чтобы выяснить, изменения каких параметров оказывали наибольшее влияние на предсказанный размер вспышки. Таким образом можно отобрать наиболее важные показатели, вмешательства по изменению которых будут являться наиболее эффективными.

В целом детерминистские модели дают достаточно возможностей для анализа и прогнозирования развития эпидемического процесса, однако они не могут ответить на один вопрос — почему в одних популяциях с одним и тем же поведением одно и то же заболевание либо генерирует вспышку, либо нет. Для анализа подобных случаев необходимо стохастическое моделирование²⁸.

²⁸ Оно, правда, тоже не ответит на вопрос, *почему* так происходит, но сможет продемонстрировать, что такое развитие событий возможно.

3.2. Стохастические модели

Стохастические модели являются вероятностными. Они не базируются на предположении о действии закона больших чисел и поэтому больше адаптированы для анализа возможного течения эпидемий в небольших популяциях²⁹.

Прежде чем начинать анализ течения эпидемии в рамках стохастической модели, стоит разобраться в некоторых предположениях, которые лежат в их основе [14]. Предположим, что нашей целью является выяснить, какое количество здоровых людей имели контакт с заразным пациентом в течение определенного периода времени и заразились. Если известно, что в среднем каждый заразный пациент имеет с потенциально заразных контактов в день, то вероятность того, что определенный контакт будет заразным, определяется биномиальным распределением с параметром p , равным:

$$p = \frac{c}{I + S - 1},$$

где I — количество заразных пациентов, а S — количество здоровых.

В том случае, если общий размер популяции достаточно большой, то в знаменателе приведенной выше формулы можно просто поставить суммарную численность популяции $I+S$.

В том случае, если вероятности контакта между заразным и здоровым членами популяции являются одинаковыми (предположение гомогенности скрещивания) и вероятность заразиться в случае контакта составляет β , то очевидно, что вероятность контакта, при котором произойдет заражение, составляет $\beta * p$. Тогда вероятность не заразиться, если в популяции присутствует один инфицированный, составляет $(1 - \beta * p)$. Если это так, то вероятность не заразиться в течение данного периода времени (дня) равна произведению вероятностей не заразиться от каждого из I заразных пациентов в популяции, т.е. она равна $(1 - \beta * p)^I$. Соответственно, вероятность заразиться в течение данного периода времени, π , равна

$$\pi = 1 - (1 - \beta * p)^I.$$

Количество заразившихся в течение данного периода (дня) неизвестно, однако оно подчиняется биномиальному распределению вероятностей:

$$P(k) = \binom{S}{k} \pi^k (1 - \pi)^{S-k}$$

Соответственно среднее количество заражений μ составит:

$$\mu = S * \pi,$$

а стандартное отклонение для количества заражений σ будет равно:

$$\sigma = \sqrt{S * \pi * (1 - \pi)}$$

Можно попытаться аппроксимировать π , если разложить формулу по правилам разложения бинома Ньютона и взять только первые два члена. Тогда можно получить:

²⁹ Хотя могут использоваться и для анализа и прогноза развития ситуации и в больших популяциях.

$$\pi = 1 - \left[1 - \beta * I * p + \binom{I}{2} \beta * p - \dots \right]$$

$$\pi \approx 1 - (1 - \beta * I * p) = \beta * I * p = \frac{\beta * c * I}{I + S}$$

Соответственно, среднее количество заражений составит

$$\mu = \frac{\beta * c * I * S}{I + S},$$

а стандартное отклонение

$$\sigma = \sqrt{\frac{\beta * c * I * S}{I + S} \left(1 - \frac{\beta * c * I}{I + S} \right)}$$

Размер стандартного отклонения по отношению к среднему часто используется как показатель того, насколько вероятны сильные отклонения от среднего. Если это отношение достаточно мало, то отклонения от среднего будут небольшими, и можно достаточно спокойно игнорировать стохастический характер эпидемии (поскольку тогда среднее количество зараженных в день будет практически всегда равно μ). Однако, если это отношение достаточно велико, это означает, что реальное количество зараженных в день будет меняться от одного дня к другому и игнорировать стохастический характер эпидемии является неправильным. Соответственно, можно проанализировать, от чего зависит отношение стандартного отклонения к среднему, называемое коэффициентом вариации:

$$CV = \frac{\sigma}{\mu} = \sqrt{\frac{S * \pi * (1 - \pi)}{S * \pi}} = \sqrt{\frac{(1 - \pi)}{S * \pi}}$$

Раскрывая значение π ,

$$CV = \frac{\sigma}{\mu} = \sqrt{\frac{(I * \beta * c - I - S)}{c * \beta * I * S}} = \sqrt{\frac{(I * (\beta * c - 1) - S)}{c * \beta * I * S}}$$

Если обозначить произведение вероятности заразиться в случае контакта β и количества контактов c как $\lambda = \beta * c$, и принять процент инфицированных в популяции равным θ ($I = \theta * N$, $S = (1 - \theta) N$), то выражение для коэффициента вариации становится следующим:

$$CV = \frac{\sigma}{\mu} = \sqrt{\frac{(I * (\lambda - 1) - S)}{\lambda * I * S}} = \sqrt{\frac{1 - \theta * \lambda}{N * \lambda * \theta (1 - \theta)}}$$

Из этого выражения видно, что, чем больше зараженных в популяции и чем больше заразность заболевания и количество контактов (λ), тем больше будет коэффициент вариации и меньше причин для использования детерминистской модели развития инфекционного заболевания. С другой стороны, чем больше размер популяции, тем меньше будет коэффициент вариации, и детерминистские модели будут более обоснованными. Для иллюстрации приведем таблицу (табл. 6), иллюстрирующую значения коэффициента вариации в зависимости от заразности заболевания (предположим, что количество контактов c является одинаковым и составляет 4 в единицу времени).

Эта таблица четко показывает, что в больших популяциях использование детерминистских моделей не ведет к значительным ошибкам в предсказании (в табл. 6 при размерах популяции в пять тысяч человек ошибка даже при заразности в 10% не превышает 7%). Вместе с тем для малозаразных заболеваний, таких, как, например, ВИЧ-инфекция, детерминистские модели смогут быть относительно справедливыми только при длительных периодах наблюдения (как было показано выше, вероятность заражения зависит от количества половых актов с партнером в единицу времени и, соответственно, чем длиннее период времени, тем вероятность заражения выше) и в крупных популяциях³⁰.

Таблица 6

Зависимость коэффициента вариации от заразности заболевания и размера популяции

β	θ	N	CV
0,1	0,1	50	0,73
0,5	0,1	50	0,30
0,7	0,1	50	0,24
0,1	0,1	500	0,23
0,5	0,1	500	0,09
0,7	0,1	500	0,08
0,1	0,1	5000	0,07
0,5	0,1	5000	0,03
0,7	0,1	5000	0,02

Поскольку количество зараженных в день в приведенных выше рассуждениях являлось случайной величиной, то и количество зараженных в последующие дни (эпидемическая кривая) будет также случайной величиной. В противоположность детерминистской модели мы уже не можем говорить о течении эпидемии, а лишь о распределении возможных эпидемических кривых.

Можно показать, что для любого небольшого промежутка времени Δt ,

$$\pi(\Delta t) = \frac{\lambda}{N} * I * \Delta t$$

т.е. вероятность того, что данный здоровый человек заразится в результате контакта с одним из I зараженных за период времени, равный Δt , пропорциональна количеству заразных лиц в популяции, заразности заболевания и длительности периода наблюдения. Эта вероятность, π , относится к каждому из здоровых лиц в популяции (при допущении гомогенности контактов), соответственно вероятность заразиться за период Δt равна π , а вероятность не заразиться равна $1 - \pi$. Поскольку количество здоровых в популяции равно S , то вероятности заражения определенного количества пациентов (при малых промежутках времени) определяются так:

- Вероятность отсутствия заражения: $(1 - \frac{\lambda}{N} * I * \Delta t)^S \approx 1 - \frac{1}{N} * S * I * \Delta t$
- Вероятность заражения одного пациента: $\binom{S}{1} \frac{\lambda}{N} * I * \Delta t * (1 - \frac{\lambda}{N} * I * \Delta t)^{S-1} \approx \frac{\lambda}{N} * I * \Delta t$
- Вероятность заражения более одного пациента равна нулю (очень малая величина).

³⁰ В случае ВИЧ-инфекции, если период наблюдения составляет год и количество контактов в течение года 104, то ошибка менее 3% начинает наблюдаться только при размере популяции не менее 100 000 человек.

До настоящего момента обсуждался только вопрос о заражении пациентов. Однако большинство инфекционных заболеваний имеют ограниченный заразный период либо ввиду того, что иммунная система пациента справляется с возбудителем, либо за счет того, что пациент погибает. При этом, если иммунная система справляется с возбудителем, возможны два варианта — пациент получает длительный иммунитет к возбудителям этого заболевания, либо, выздоровев, он может снова заразиться³¹. Если после выздоровления пациент может заразиться снова, такие заболевания описываются SIS-моделями. Если пациент выздоравливает с формированием длительного иммунитета или же он погибает, т.е. удаляется из пула восприимчивых к инфекции, то такие инфекции описываются SIR-моделями. Напомним, что истинные SIR-модели в открытых популяциях, где снижение количества восприимчивых компенсируется притоком новых членов в популяцию (за счет миграции или рождения) могут быть сведены к SIS-моделям.

Поэтому мы продолжим рассмотрение стохастических моделей вариантом SIS-модели. В этой модели после выздоровления пациент снова становится уязвимым и, соответственно, общая сумма инфицированных и уязвимых пациентов постоянна и равна численности популяции.

Вероятность выздоровления для каждого конкретного пациента зависит только от продолжительности болезни. При этом могут быть различные функциональные формы, описывающие вероятность выздоровления как функцию времени с момента заражения. Для простоты можно принять, что вероятность выздоровления обратно пропорциональна средней продолжительности болезни. Эта достаточно очевидная зависимость свидетельствует об экспоненциальном распределении времен продолжительности болезни и часто используется в моделировании инфекционных (и других) заболеваний³².

Достаточно легко показать, что, если вероятность выздоровления определяется величиной $\frac{1}{D}$, где D — средняя продолжительность заболевания, то в момент $t + \Delta t$ ³³ распределение количества выздоровевших подчиняется следующим закономерностям:

- Никто не выздоровел: $1 - \frac{I}{D} \Delta t$
- Выздоровел ровно 1 человек: $\frac{I}{D} \Delta t$
- Выздоровело более 1 человека: 0

Теперь для построения модели рассмотрим распределение времен до наступления изменений в популяции. Под изменениями будем понимать заражение или выздоровление человека. Очевидно, что распределение этих времен является случайным. Каждый раз, когда такой момент наступает, можно задать вопрос — а какое изменение произойдет в популяции? Вероятность заражения, как уже описывалось выше, определяется вероятностью $\frac{\lambda}{N} * I * S * \Delta t$, а выздоровление — $\frac{I}{D} * \Delta t$. Соответственно, поскольку рассматривается момент, в который какое-то событие уже произошло, вероятность того, что этим событием было заражение, составит:

$$\frac{\frac{\lambda}{N} * I * S * \Delta t}{\frac{\lambda}{N} * I * S * \Delta t + \frac{I}{D} * \Delta t} = \frac{\frac{\lambda}{N} * I * S}{\frac{\lambda}{N} * I * S + \frac{I}{D}},$$

³¹ Это, конечно, упрощение, иммунитет может ослабевать по мере прохождения времени, а повторные заражения, например в случае *N.gonorrhoeae* вызваны выраженными антигенными различиями между разными штаммами этого возбудителя, на каждый из которых формируется обычный иммунный ответ.

³² В частности, марковские модели предполагают именно экспоненциальное распределение времен пребывания в каждой из болезненных стадий.

³³ Если Δt достаточно мало.

а того, что этим событием станет выздоровление:

$$\frac{\frac{I}{D} * \Delta t}{\frac{\lambda}{N} * I * S * \Delta t + \frac{I}{D} * \Delta t} = \frac{\frac{I}{D}}{\frac{\lambda}{N} * I * S + \frac{I}{D}}$$

Теперь следует ответить на вопрос о том, как часто происходят эти события, или каково распределение времен наступления событий (заражение или выздоровление).

Оказывается, что легче всего описать время до наступления события экспоненциальной зависимостью. Именно это было показано ранее, в разделе, посвященном детерминистским моделям. Таким образом, время до наступления события (либо заражения, либо выздоровления) определяется экспоненциальной зависимостью с параметром, равным величине, обратной сумме вероятностей заражения и выздоровления³⁴:

$$\frac{\lambda}{N} * I * S + \frac{I}{D}$$

Теперь все необходимые параметры и взаимоотношения между ними определены и можно приступать к формулировке модели. Идея, которая заложена в формулировке модели, заключается в том, что мы имитируем взятие случайных наблюдений из популяции возможных времен заражения и выздоровления и на основании этого реконструируем течение эпидемии.

На практике построение модели сводится к тому, что вначале определяется момент времени, который будет рассматриваться (момент наступления события). Как было указано выше, время до наступления события распределено экспоненциально с параметром, обратным сумме показателей вероятностей инфицирования и выздоровления. Для того, чтобы определить время конкретного моделируемого события, надо взять случайным образом значение из этого распределения. Для этого можно взять случайную величину, равную:

$$W_i = \frac{-\ln(Rand)}{\frac{I}{D} + \frac{\lambda}{N} * I * S},$$

где Rand — случайная величина, равномерно распределенная на отрезке 0;1.

Зная время наступления события, можно определить, каков был характер этого события. Выше уже были показаны вероятности того, что это событие будет выздоровлением или инфицированием. Для того, чтобы смоделировать ситуацию с конкретным событием, можно взять случайную величину, равномерно распределенную на отрезке, и сравнить ее с вероятностью, например, инфицирования. Если полученная величина будет меньше нее, то можно считать, что наступило инфицирование, если больше — выздоровление.

Подобный подход является возможным, поскольку случайная величина распределена равномерно и, значит, при многократной генерации этих случайных чисел их количество, оказавшееся меньше произвольно выбранной величины, находящейся в пределах [0;1], будет равно этой величине. В нашем случае это означает, что при многократном повторении процедуры мы получим именно то соотношение заражений и выздоровлений, которое соответствует приведенным выше вероятностям.

³⁴ т.е. чем она выше, тем больше событий происходит на ранних сроках

После проведения достаточно большого количества подобных имитаций ситуации наступления событий у аналитика имеется материал, соответствующий возможной траектории развития эпидемии. Повторение описанного алгоритма с начала позволит проанализировать другую траекторию эпидемии, затем еще одну и т.д. В результате аналитик может получить набор траекторий развития эпидемии для данных параметров модели и определить, какие траектории более вероятны, а какие — нет. Следует, однако, отметить, что в стохастическом моделировании (как, впрочем, и в детерминистском) нет возможности сказать, какая траектория реализуется в данном конкретном случае, однако описать наиболее вероятную — можно.

Теперь проиллюстрируем эти теоретические положения на примере создания и анализа простейшей SIS-модели эпидемии в системе SAS. Вот как выглядит код для стохастического моделирования.

```
1. DATA SIM1;
2. x=99; y=1; lam=0.5; D=10; time=0;
3. DO WHILE ((y NE 0) AND (time<500));
4.     w1 = -LOG(RANUNI(12345))/(y/D + lam*x*y/(x+y));
5.     time = time + w1;
6.     rand = RANUNI(23456);
7.     check = ((lam/(x+y))*x*y)/(y/D + (lam/(x+y))*x*y);
8.     IF rand LE check THEN DO;
9.         x=x-1; y=y+1; END;
10.    IF rand GT check THEN DO;
11.        y=y-1;x=x+1;;
12.    END;
13.    sim1 = y;
14.    OUTPUT;
15. END;
16. KEEP sim1 time;
17. RUN;
```

Программа начинается с того, что устанавливаются параметры для моделируемой популяции. В данном случае численность популяции составляет 100 человек, из которых в начальный момент времени ($time = 0$) 99 человек являются здоровыми (или уязвимыми, $x = 99$), а заражен инфекционным заболеванием один человек ($y = 1$). Количество заразных контактов в единицу времени (произведение числа контактов на вероятность заражения в каждом из них $\lambda = \beta * c$) составляет 0,5 ($lam = 0,5$), а продолжительность болезни составляет десять базовых единиц времени (например, дней, $D = 10$)³⁵. После того, как базовые параметры установлены, начинается собственно процесс моделирования. Вначале определяется, когда произойдет следующее событие. Определяется это на основании параметров экспоненциального распределения, как описано выше. Следует обратить внимание, что время до следующего события ($w1$) зависит от показателей вероятности заражения и выздоровления, но при этом является случайной величиной, о чем свидетельствует ее зависимость от случайной величины, генерируемой функцией RANUNI(). Надо помнить, что практически во всех компьютерных программах функции генерации случайных чисел генерируют в реальности последовательности псевдослучайных чисел, и эти последовательности являются одинаковыми, если известно стартовое значение. Стартовым

³⁵ Это те же параметры, что ранее использовались для описания эпидемической кривой при помощи детерминистских моделей (см. рис. 10 и рис. 11).

значением является аргумент функции RANUNI. Если этот аргумент равен 0 (или меньше), то каждый раз генерируются разные последовательности (для инициализации используются показатели счетчика времени), однако в данном случае использовано положительное ненулевое значение для того, чтобы получающаяся эпидемическая кривая была предсказуемой. Для получения серии эпидемических кривых аргумент функции RANUNI должен быть нулевым.

После определения времени наступления следующего события оно переводится в показатель времени от начала вспышки (time) и затем определяется, каким было это событие. Для этого генерируется новая случайная величина, которая сравнивается с пограничным значением, определяющим вероятности того, что это событие было заражением или выздоровлением. Если решается, что событием было заражение, количество уязвимых уменьшается на единицу, а количество инфицированных увеличивается на единицу. Если событием было выздоровление — ситуация меняется на обратную — количество инфицированных уменьшается на единицу, а количество здоровых восприимчивых пациентов увеличивается на ту же единицу.

Процесс повторяется до тех пор, пока в популяции остается хотя бы один инфицированный или же эпидемия продолжается на протяжении более 100 временных единиц. Если затем полученные данные изобразить графически, результат будет следующим:

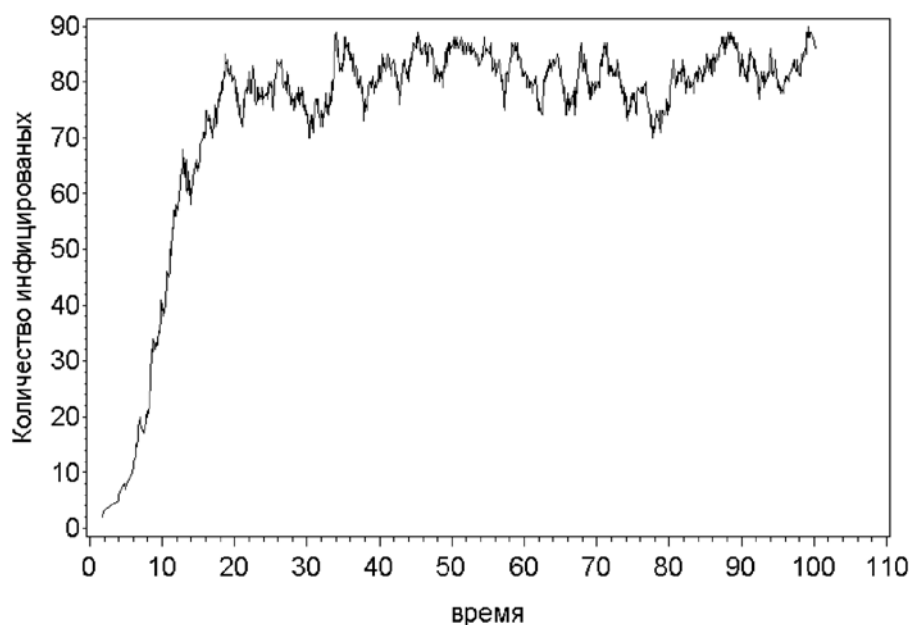


Рис. 19. Количество инфицированных, предсказанное при помощи стохастической модели для тех же параметров, что и на рис. 10 и рис. 11.

Как видно из рис. 19, количество инфицированных быстро возрастает до уровня примерно 80 человек) из 100 (т.е. примерно до 80%), и затем достаточно выражено колеблется вокруг этого значения, то повышаясь (иногда почти до 90% популяции), то снижаясь (иногда до 70% популяции). Сравнение с детерминистским моделированием (рис. 10 и рис. 11) показывает, что распространенность заболевания в случае использования стохастической модели практически такая же, как и при использовании детерминистской, однако колебания распространенности значительно выше при стохастической модели.

Теперь попробуем разобрать немного более сложный, но более реальный пример. Предположим, что есть желание смоделировать распространение гонореи в популяции небольшого города после

³⁶ При одном случае и описываемых ниже стартовых параметрах эпидемия вспышка может вообще не развиваться.

внесения туда двух завозных случаев³⁶. Для моделирования нам понадобятся данные уже (как минимум) для двух групп — мужчин и женщин.

Предположим, что результаты опроса населения репродуктивного возраста продемонстрировали, что в городе имеется около 1000 мужчин, которые не состоят в моногамных отношениях. Они сообщили, что в месяц у них бывает в среднем 2 сексуальных партнера. Вероятность заражения гонореей от одного партнера составляет 50%³⁷. Продолжительность заразного периода при более или менее своевременном выявлении и лечении составляет 1,8 месяца, без адекватного выявления и лечения — полгода [21]. Предполагая смешанную ситуацию в данном городе, будем считать, что продолжительность заразного периода составляет 3 месяца для мужчин и 4 месяца для женщин. Опрос выборки женщин, которые заявили, что не состоят в моногамных отношениях, показал, что они сообщили о 6 половых партнерах в месяц (среди этих женщин могли быть, например, работницы коммерческого секса). Численность этой группы установить трудно, однако, зная численность мужчин и соотношение количества партнеров, его можно оценить по формуле $N_f = \frac{C_m}{C_f} N_m$, где индексы m и f обозначают показатели для мужчин и женщин соответственно. Располагая этими данными, программа стохастического моделирования для данного города будет выглядеть так:

```

1.   %LET l_model=36;
2.   DATA SIM1;
3.   Xm=999; Ym=3; Cm=2; beta=0.5; LAMm=Cm*beta;
4.   Dm=1.8; TIMEm=0; Nm=Xm+Ym;
5.   Cf=6; Nf=Cm/Cf*Nm; Yf=0; Xf=Nf-Yf;
6.   LAMf=Cf*beta; Df=4; TIMEf=0;
7.   DO WHILE ((Yf > 0 OR Ym > 0)
8.           AND (TIMEm<&l_model)
9.           AND (TIMEf<&l_model));
10.  rand = RANUNI(12345);
11.  * мужчины;
12.  Sm = -LOG(RANUNI(1356))/(Ym/Dm + LAMf*Yf*Xm/(Nm));
13.  TIMEm = TIMEm + Sm;
14.  check = (LAMf*Yf*(Xm/Nm))/(Ym/Dm + LAMf*Yf*(Xm/Nm));
15.  IF rand LE check THEN DO;
16.      Xm=Xm-1; Ym=Ym+1; ***Заражение***; END;
17.  IF rand GT check THEN DO;
18.      Ym=Ym-1; Xm=Xm+1; ***Выздоровление***;
19.  END;
20.  If Ym<0 then Ym=0;
21.  SIMm = Ym/Nm;
22.  * женщины;
23.  Sf = -LOG(RANUNI(2356))/(Yf/Df + LAMm*Xf*Ym/(Nf));
24.  TIMEf = TIMEf + Sf;
25.  check = (LAMm*Ym*(Xf/Nf))/(Yf/Df + LAMm*Ym*(Xf/Nf));
26.  IF rand LE check THEN DO;

```

³⁷ На самом деле это вероятность заражения за один половой акт, соответственно, если партнерство включает более одного полового акта, вероятность заражения выше.

```

27.           Xf=Xf-1; Yf=Yf+1; ***Заражение***; END;
28.           IF rand GT check THEN DO;
29.             Yf=Yf-1;Xf=Xf+1; ***Выздоровление***;
30.           END;
31.           If Yf<0 then Yf=0;
32.           SIMf = Yf/Nf;
33.           OUTPUT;
34.         END;
35.         KEEP SIMm SIMf TIMEm TIMEf;
36.         RUN;

```

Сама модель практически аналогична стохастической модели, представленной ранее, однако в ней было необходимо учесть наличие двух групп. Понятно, что мужчины могут заразиться только от женщин и наоборот. Соответственно, для мужчин необходимо основывать расчеты на количестве зараженных женщин и наоборот. Поэтому в формулах для вероятностей заражения у мужчин стоит количество контактов женщин, в результате которых должна была бы произойти передача инфекции, а в формулах для женщин — соответствующие показатели, взятые из мужской группы. Для того, чтобы результаты было проще интерпретировать, они приводятся в процентном отношении к численности группы в целом. Результаты моделирования на период, равный 36 месяцам, представлены на рис. 20.

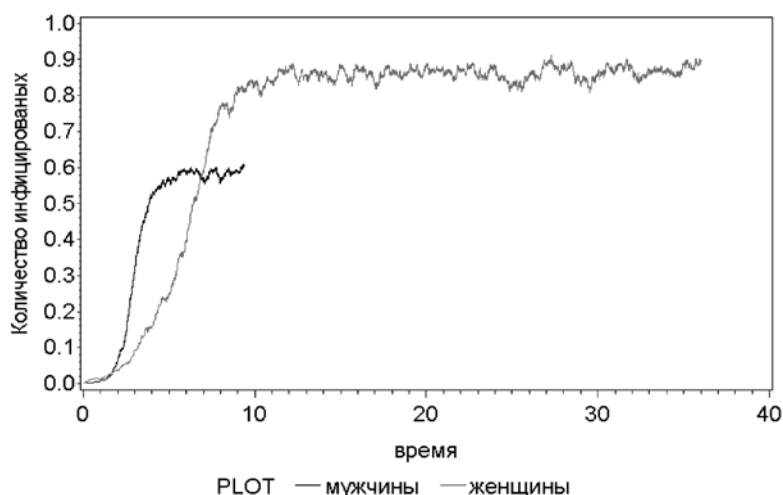


Рис. 20. Стохастическое моделирование распространения эпидемии в двух взаимодействующих группах людей.

Как видно из этого рисунка, пораженность женщин нарастает достаточно резко и стабилизируется на высоком уровне с относительно небольшими колебаниями. Для мужчин рост более быстрый, хотя достигнутый уровень пораженности немного ниже. Надо отметить, что полученные результаты не сильно отличаются от таковых в случае, если бы использовалась детерминистская модель (это во многом связано с достаточно высоким репродуктивным числом для инфекции в данной гипотетической популяции: $R_0 = \beta * c * D = 0,5 * 2 * 1,8 = 1,8$ для мужчин и $R_0 = 0,5 * 6 * 4 = 12$ для женщин). Очевидно, что женщины являются в данной ситуации «двигателем» эпидемии. Если использовать более реальные цифры для количества партнеров (1,28 для женщин и 2,47 для мужчин в год [5] — 0,11 и 0,206 в месяц, соответственно), то никакой эпидемии не наблюдается. Даже если брать значения для группы лиц с венерическими заболеваниями [5] (0,57 в месяц для женщин и 0,36 для мужчин) модель предсказывает угасание эпидемии после различных периодов — от одного до 18 месяцев³⁸. Этот вывод можно было бы сделать и на осно-

вании анализа репродуктивного числа, которое в последнем случае составляет 1,14 для женщин и 0,32 для мужчин.

Возникает вопрос, а зачем все-таки нужны стохастические модели, если они не могут достаточно точно предсказать течение эпидемии, если размер популяции мал или вероятность заражения невелика, а если моделируется распространение заболевания в большой популяции (или заразность относительно велика), то детерминистские модели дают ничуть не худший прогноз, только с значительно меньшими проблемами³⁹?

Ответ на самом деле достаточно простой: иногда исследователя интересует не наиболее вероятный курс эпидемии, а вероятность определенной ситуации. Например, из обсуждения репродуктивного числа, которое было сделано выше, ясно, что если оно ниже единицы, вспышки быть не должно. Более того, детерминистская модель предполагает, что в таком случае эпидемическая кривая затухает по экспоненте и количество случаев во второй и третьей волнах эпидемии должно быть меньше, чем в первой. Иными словами, если заболевание мало заразное и в популяции всего один больной, вспышки не будет. Точка. Дело закрыто.

Следуя этой логике, не удивительно, что ряд исследователей, в частности D.Gisselquist, поставили под сомнение оценки вероятности внутрибольничного заражения ВИЧ-инфекцией [10], поскольку имевшиеся в литературе результаты оценки вероятности заражения при случайном уколе иглой [20] указывали на крайне низкую заразность ВИЧ. Вероятность заражения составляла 0,27%. И никакая детерминистская модель при более-менее разумном числе инъекций в больнице не могла бы предсказать возникновение вспышки, например, такой, как наблюдалась в больнице Эль Фатех в Бенгази (Ливия)⁴⁰. С другой стороны, количество столь крупных внутрибольничных вспышек достаточно мало и, соответственно, если правы Gisselquist и соавт. [10], то почему больше таких вспышек не наблюдалось? Неужели все остальные больницы во всем мире являются идеалом соблюдения противоэпидемических мероприятий? Вряд ли.

Таким образом, получается, что вспышка в Бенгази была в некотором смысле aberrацией, случайностью. И вот тут-то и возникает надобность в стохастических моделях, которые как раз могут предсказать вероятность возникновения подобных необычных вспышек.

Попробуем оценить вероятность возникновения подобной вспышки на основании имеющейся информации. Будем использовать SIS-модель, поскольку обычно больницы стараются, чтобы койки не пустовали и, соответственно, численность популяции в больнице является стационарной и выбытие (в результате смерти или перевода в другую больницу) тут же компенсируется новой госпитализацией. Предположим, что вероятность заражения при однократном уколе иглой составляет 0,3%, количество уколов в течение дня составляло 7, и пациент находится в стационаре 30 дней (эти значения примерно равны данным для детей, заразившихся ВИЧ-инфекцией в результате вспышки в Элисте, Калмыкия [6]). Если использовать для анализа детерминистский подход то репродуктивное число оценивается следующей величиной: $R_0 = 7 * 30 * 0,003 = 0,63$, т.е. оно значительно меньше единицы и вспышка невозможна. Посмотрим же, что скажет на это стохастическая модель. Сама программа является простой модификацией стохастической SIS-модели, описанной выше.

³⁸ Это означает, что в популяции в реальности есть небольшие группы с большим количеством контактов, которые поддерживают существование подобных заразных инфекций в популяции.

³⁹ Собственно, именно поэтому детерминистские модели используются значительно чаще в реальных приложениях. При этом стоит помнить, что кроме популяционных детерминистских моделей существуют индивидуальные модели, авторы которых пытаются смоделировать все взаимодействия между людьми, наиболее известным примером является EpiSIMS[2], модель распространения оспы в Портленде после террористической атаки.

⁴⁰ Там ВИЧ-инфекцией заразились более 400 детей [16, 15].


```

1. DATA BASELINE;
2. beta=0.003;
3. c=7;
4. lam=beta*c;
5. D=30;
6. x=500; y=1; time=0; Infected=0;
7. OUTPUT;
8. RUN;
9. DATA SIM1;
10. SET BASELINE;
11. Kwa=0;
12. DO WHILE ((y NE 0) AND (Kwa<500));
13.     s = -LOG(RANUNI(0))/(y/D + lam*x*y/(x+y));
14.     time = time + s;
15.     Kwa+1;
16.     rand = RANUNI(0);
17.     check = ((lam/(x+y))*x*y)/(y/D + (lam/(x+y))*x*y);
18.     IF rand LE check THEN DO;
19.         x=x-1; y=y+1; Infected=Infected+1; END;
20.     IF rand GT check THEN DO;
21.         y=y-1; x=x+1;
22.     END;
23. END;
24. sim1 = Infected;
25. OUTPUT;
26. KEEP sim1 time;
27. RUN;
28. DATA sim1;
29. SET sim1;
30.     ki=&i;
31. RUN;
32. %IF &i=1 %THEN %DO;
33.     DATA mum1;
34.         SET sim1;
35.         RUN;
36. %END;
37. %ELSE
38. %DO;
39. PROC APPEND BASE=mum1 DATA=sim1 FORCE;
40. RUN;
41. %END;
42. %END;

```

- 43. %MEND;
- 44.
- 45. %stoch;

Модель учитывает количество инфицированных за время вспышки (параметр K_{Wa}) при повторении модели 10 000 раз. В результате запуска этой программы были получены следующие результаты⁴¹:

- В большинстве случаев (61,2%) никакой вспышки не будет
- В 14,9% заразится 1 человек
- В 6,7% случаев — 2 человека
- В 4,2% случаев — 3 человека
- В 9,3% случаев во вспышке заразится более 10 человек, причем возможны вспышки с 50 и более зараженными (вероятность вспышки с более чем 50 зараженными 0,3%)

Таким образом, стохастическая модель показывает возможность развития достаточно значительных вспышек там, где детерминистская модель предполагала отсутствие таковых. Более того, если попытаться проанализировать вероятность заражения, опираясь только на одну вспышку (с наибольшим количеством зараженных), вероятность заражения при однократном контакте может оказаться значительно большей, чем она есть на самом деле. Поэтому в случае редких событий стохастические модели являются единственно приемлемым методом изучения и прогнозирования ситуации.

Еще одним приложением стохастических моделей является оценка продолжительности вспышки в случае, если репродуктивное число меньше единицы (либо если используется SIR-модель). В частности, для описанного выше примера с популяцией лиц, страдающих венерическими болезнями (когда суммарное репродуктивное число меньше единицы⁴²), моделирование показывает, то в 20% случаев вспышки не будет, в 45% случаев вспышка продлится до полугода, еще в 17% случаев до года, но примерно в 0,2% случаев возможно продолжение вспышки от шести лет и более. Более того, рассмотрение эпидемических кривых с длительными периодами существования инфекции в популяции показывает несколько подъемов и падений. Этот пример демонстрирует дополнительные возможности стохастического моделирования при описании эпидемического процесса и то, что наблюдаемые изменения заболеваемости и распространенности могут иногда являться следствием не изменений поведения популяции или социальных условий, а просто результатом случайных процессов⁴³.

Вместе с тем, из приведенных выше рассуждений очевидно, что для адекватного моделирования инфекционного процесса необходимо знать не только вероятности заражения, но и обладать информацией о популяции риска — ее размерах и особенностях взаимодействия членов популяции друг с другом. В случае воздушно-капельных инфекций это относительно просто, поскольку можно считать, что скрещивание происходит практически случайно для всего населения⁴⁴.

⁴¹ Если повторять запуски программы, результаты могут быть немного отличными, но при большом числе повторений картина будет практически одинаковой.

⁴² Напомним, количество смен партнеров в месяц 0,57 в месяц для женщин и 0,36 для мужчин.

⁴³ Правда, как отмечалось неоднократно выше, размер популяции должен быть относительно небольшим. В большой популяции эффект случайных факторов достаточно быстро нивелируется.

⁴⁴ Это на самом деле упрощение, ряд людей — врачи, учителя, продавцы — имеют больше контактов, чем другие. Реальная модель, как, например, EpiSIMS [2], должна учитывать эти взаимодействия, так же как и тот факт, что люди, живущие ближе друг к другу в городе, встречаются чаще (магазины, станции метро и т.п.).

Однако, если речь заходит о заболеваниях, передающихся преимущественно половым путем, или гемоконтактных инфекциях, то там группы риска часто весьма сильно отличаются по численности от численности населения. Соответственно, для адекватного моделирования необходимо располагать оценкой численности этих групп риска, например, потребителей инъекционных наркотиков или работниц коммерческого секса и их клиентов.

Аналізу численности трудно наблюдаемых популяций посвящены методики, объединяемые названием методик двойного охвата.

4. Методы двойного охвата

4.1. Метод Линкольна-Петерсена

Методы двойного охвата появились в биологии, когда возникла необходимость в изучении размеров популяции животных. Было очевидно, что зайцы и лисы вряд ли будут выстраиваться в очередь или сидеть в норах, ожидая прихода сотрудника службы переписи населения. Они, наоборот, будут прятаться и избегать контакта с исследователями. Чтобы разрешить возникшую проблему, был предложен простой вариант — надо попытаться отловить группу животных и пометить их (окольцевать), а затем отпустить на волю. Затем снова отловить группу животных и подсчитать среди них количество окольцованных. Понятно, что, чем больше в лесу животных, тем меньше их будет во второй выборке (они будут разбавлены не окольцованными). Если же в лесу животных мало, то количество окольцованных во второй выборке будет большим. В крайнем случае, когда в первый раз отловили всех животных, живущих в данном лесу, во второй выборке все животные будут окольцованными. Соответственно, численность популяции в целом обратно пропорциональна проценту окольцованных во второй выборке. Если x_1 — численность лиц в первой выборке, то она может быть разделена на две подгруппы — x_{12} , субъекты, которые были в первой и второй выборках, и x_{10} — которые были только в первой выборке. Численность лиц во второй выборке x_2 тогда равна сумме x_{12} и x_{02} — лиц, которые попали только во вторую выборку. Соответственно, процент лиц во второй выборке, попавших и в первую выборку

$$p = \frac{x_{12}}{x_2}$$

Опираясь на описанное выше соотношение для общей численности популяции, можно записать:

$$N = \frac{x_1}{p} = \frac{x_1 * x_2}{x_{12}}$$

Эта формула называется формулой Линкольна⁴⁵-Петерсена⁴⁶ и является самой простой из методов двойного охвата⁴⁷. Можно проиллюстрировать ее использование так — если в первую выборку попало 20 субъектов, из которых во второй выборке из 30 субъектов оказались 10, то общая численность популяции составляет

$$N = \frac{20 * 30}{10} = 60$$

Конечно, возникает вопрос о точности полученных оценок. Очевидно, что чем меньше количество лиц в первой и второй группах, тем ниже будет точность оценки, а чем больше выборки — тем она точнее. Соответственно можно предполагать, что показатель точности — дисперсию — можно оценить на основании количества включенных в выборки лиц:

$$Var_N = \frac{x_1 * x_2 * (x_1 - x_{12}) * (x_2 - x_{12})}{x_{12}^3}$$

⁴⁵ Фредерик Линкольн был сотрудником службы рыбного и лесного хозяйства США начиная с 20-х годов XX века и разработал методику оценки размеров популяции птиц при помощи их кольцевания.

⁴⁶ Карл Георг Йоханнес Петерсен являлся директором датской биологической станции. Он изобрел медную метку, которую прикреплял рыбам для изучения их миграции. Когда треть меченых рыб попала рыбакам, он сообразил, что эти данные можно использовать для оценки размера популяции рыб. Его результаты были опубликованы в 1896 году.

⁴⁷ На самом деле первые публикации на эту тему были у Пьера Симона Лапласа в XVII столетии.

В том случае, если количество наблюдений в ячейках менее 50, лучше использовать формулы с коррекцией. Для размера популяции:

$$N = \frac{(x_1 + 1) * (x_2 + 1)}{x_{12} + 1}$$

И для дисперсии:

$$Var_N = \frac{(x_1 + 1) * (x_2 + 1) * (x_1 - x_{12}) * (x_2 - x_{12})}{x_{12}^2 * (x_{12} + 2)}$$

Тогда, для приведенного выше примера, откорректированные значения численности популяции составляют 59, а дисперсии — 108,5. На основании рассчитанной дисперсии можно оценить доверительный интервал для оценки численности популяции.

$$95\%CI = N \pm 1.96 * \sqrt{Var_N}$$

Для приведенного выше примера ширина 95% доверительного интервала составит 41, а сам интервал будет равен 39–80 единиц. Иными словами, реальная численность популяции составляет где-то между 40 и 80 индивидами. Полученная величина показывает, какова в реальности может быть численность популяции и, соответственно, задачей исследователей является получение по возможности более точных оценок.

В эпидемиологии метод Линкольна-Петерсена базируется на получении информации из нескольких источников. Так, например, для определения численности потребителей инъекционных наркотиков в каком-то регионе можно воспользоваться такими источниками, как данные учета лиц, страдающих наркоманией, в наркологическом диспансере и данные госпитализации лиц с передозировкой наркотиков⁴⁸.

Приведем достаточно реальный пример использования метода двойного охвата с оценками по Линкольну-Петерсену. Для определения численности потребителей инъекционных наркотиков (ПИН) в столице Таиланда Бангкоке исследователи получили доступ к двум источникам информации — спискам употреблявших опиаты лиц, включенных в программы лечения метадонном, в наркологических клиниках Бангкока (список 1), и к списку лиц, арестованных бангкокской полицией, проверка которых на опиаты дала положительные результаты (список 2). При сравнении этих двух списков были получены данные, приведенные в табл. 7.

Таблица 7

Присутствие ПИН Бангкока в двух списках

	Список 1	
Список 2	присутствует	нет
присутствует	171	1369
нет	3893	

Общее количество ПИН в первом списке составило 4064, а во втором — 1540. Соответственно, оценка численности ПИН в Бангкоке по методу Линкольна-Петерсена составляет 36 600 человек, а доверительный интервал для этого значения — от 31 500 до 41 700 человек. Иными словами, все-

⁴⁸ Правда, в данном случае надо бы удостовериться в том, что лица, состоящие на учете, являются активными наркоманами, иначе выборки делаются из разных популяций.

го в Бангкоке присутствует от 30 до 40 тысяч ПИН. Полученный результат можно использовать в описанных выше моделях, например, для предсказания динамики эпидемии ВИЧ-инфекции, ее размеров и потребности в профилактических мероприятиях.

Метод Линкольна-Петерсена базируется на важных допущениях — численность популяции не меняется между двумя обследованиями (популяция стационарная), процесс отбора в выборки является полностью случайным (все субъекты имеют одинаковую вероятность попадания в выборки) и вероятность попадания в первую и вторую выборки одинакова. В реальной жизни выполнить эти правила не так-то просто. В том случае, если между двумя обследованиями проходит более-менее большой промежуток времени, численность популяции может измениться — в ней могут появиться новые члены (например, количество лиц, потребляющих инъекционные формы наркотиков, может увеличиться за счет новых членов), а могут выбыть старые (например, какое-то количество ПИН может погибнуть от передозировки или быть арестовано и отправлено в места заключения). Если вероятность выбытия не зависит от того, попал ли индивид в первую выборку или нет, метод Петерсена будет оценивать численность популяции в момент взятия выборки. К чему это приведет, проиллюстрируем на примере. Предположим, что после организации первой выборки из исходной популяции выбывают 10%. Предположим, что исходная численность популяции составляла 100 человек, из которых 20 попали в первую выборку. Соответственно, в популяции после проведения первой выборки помечено 20% индивидов. В условиях 10% смертности ко времени проведения второй выборки в популяции останется 90 индивидов, 18 из которых будут помечены. Если теперь взять выборку из 20 индивидов, то помеченными будут 4 индивида (поскольку процент помеченных не изменился) и расчетная численность популяции составит $20 * 20/4 = 100$ человек, т.е. будет равна численности популяции в момент первой выборки. Если же в той же ситуации численность популяции увеличилась на 20% (до 120 человек), то пропорция лиц, присутствовавших в двух выборках, во второй выборке понизится (станет равной $20/120 = 0,17$). Соответственно, оценочная численность популяции станет равной также $20/0,17 = 120$ или будет равна численности популяции на момент второй выборки. Таким образом, при выбытии метод Петерсена оценивает размер популяции на момент взятия первой выборки, а при росте — на момент взятия второй выборки.

Вместе с тем значительно большую проблему представляют ситуации, когда вероятность попадания во вторую выборку зависит от присутствия индивида в первой (или вероятность выбытия зависит от попадания в первую выборку). Если, например, исследование проводится на основании данных по задержаниям потребителей наркотиков, очевидно, что часть после задержания будут арестованы и осуждены, а, значит, покинут популяцию. Скажем, в той же популяции из 100 человек были задержаны 20, из которых 10 затем были осуждены (те, кто не был задержан, остались в популяции, т.е. численность снизилась до 90 человек). При анализе второй группы из 20 задержанных повторное задержание будет только у 11% (сколько осталось лиц с задержаниями в популяции). Соответственно, оценочная величина популяции в момент взятия первой выборки составит $20/0,11 = 180$ человек — величина будет резко завышена. Аналогичным образом, если задержанные в первый раз чаще задерживаются и во второй (например, в два раза) — это тоже будет искажать оценки численности.

Подобные проблемы возникают и при использовании различных списков. Например, при проведении исследования по определению количества ПИН в одном городе были использованы два источника информации — списки лиц, задержанных в состоянии наркотического опьянения, и лиц, находящихся на учете в наркологическом диспансере. При анализе было установлено, что всего задержано было 2700 человек, а на учете в диспансере состояло 3506 человек, при этом в обоих списках присутствовали 1900 человек. Используя метод Линкольна-Петерсена, можно рассчитать, что численность ПИН должна составлять немногим менее 4900 человек. Однако проведенные исследования и использованием других методов и источников информации показали, что реальное количество ПИН в городе составляет около 35 тысяч человек, или

почти в семь раз больше. При детальном изучении проблемы выяснилось, что нормативными документами предписывается направление каждого задержанного в состоянии наркотического опьянения на наркологическое лечение в диспансер. Другими словами, те, кто были задержаны, имеют очень высокую вероятность попасть в группу лечения. Эти две группы положительно коррелируют и поэтому не являются независимыми.

Другой, менее очевидный, пример иллюстрирует ошибки при попытке оценить количество лиц, оказывающих коммерческие секс-услуги. Для этой оценки воспользовались двумя списками — задержанных за проституцию и данными клиники по лечению венерических заболеваний. Число задержанных составило 2700 человек, пролеченных в КВД — 3506 человек. В обоих списках оказалось 139 человек. Соответственно, оценка размера группы лиц, оказывающих коммерческие секс-услуги, составляет более 68 тысяч человек. Однако более детальный анализ показал, что их истинное число в два раза меньше. Ошибка была связана с несколькими причинами. Выяснилось, что проститутки, у которых имеется большое количество клиентов и хороший доход, задерживаются с меньшей вероятностью, поскольку они проводят большее количество времени с клиентами и меньшее количество времени на улицах, а если они задерживаются, у них есть средства на взятку и возможность откупиться до того, как их отправят в центр предварительного заключения, и они чаще обращаются в КВД, чем менее популярные проститутки. Соответственно, две выборки отрицательно коррелируют друг с другом, поэтому не являются независимыми, и оценки численности с их помощью являются ошибочными.

4.2. Методы с учетом корреляции источников

Все эти проблемы привели к попыткам разработать методологию оценки, которая могла бы корректировать возможную корреляцию между источниками и, соответственно, улучшать оценки численности популяции. Естественно, что не существует метода, который смог бы получить больший объем информации из двух источников, необходимых для метода Линкольна-Петерсена. Любая коррекция требует дополнительных источников информации в виде дополнительных списков. Однако даже наличие трех источников уже позволяет значительно улучшить возможности по получению достоверного результата.

Самым простым методом, который позволяет проанализировать, нет ли корреляции между источниками и, при необходимости, откорректировать результат, является метод Виттес [23]. Для его реализации необходимо располагать как минимум тремя источниками данных (тремя списками). При описании результатов исследования с тремя списками возможно следующее схематичное представление результатов (рис. 21).

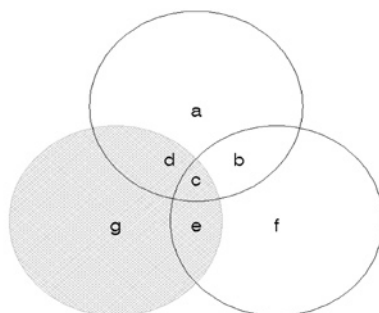


Рис. 21. Схематическое изображение результатов сравнения трех списков.

На этом рисунке каждый список представлен в виде окружности. Представители изучаемой популяции могут находиться в одном, двух или трех списках. Символы, показывающие их численность, представлены в месте пересечения окружностей. Так, например, численность

лиц, присутствующих во всех трех списках, обозначена буквой *c*, а численность лиц, которые находились только в списке, обозначенном заштрихованной окружностью, и больше ни в каком — буквой *g*. На первом этапе исследователь может провести оценку численности популяции по Линкольну-Петерсену, используя списки попарно. Очевидно, что в таком случае он получит при трех списках три значения. В том случае, если они одинаковы то это означает, что выраженной корреляции нет и оценками можно воспользоваться без дальнейшей коррекции. Если же они различаются, то необходимо выявить наиболее коррелирующие друг с другом списки.

Для этого каждый список берут в отдельности и оценивают то, как распределены в нем лица из других списков. Например, если взять на рис. 21 список, обозначенный заштрихованной окружностью, то общее количество лиц в этом списке будет равно сумме $g + d + c + e$. Показателем равномерности распределения в этом списке представителей двух других списков будет отношение шансов, равное

$$OR = \frac{c * g}{d * e}$$

Это отношение шансов показывает, нет ли связи между принадлежностью к одному и другому спискам, помеченным незаштрихованными окружностями. Такая связь может, например, выражаться в том, что те, кто оказался в одном списке, с большей вероятностью будут и в другом (положительная связь, $OR > 1$, метод Линкольна-Петерсена будет недооценивать истинную численность популяции), или же те, кто оказался в одном списке, с меньшей вероятностью окажутся в другом (отрицательная связь, $OR < 1$, метод Линкольна-Петерсена будет преувеличивать истинную численность популяции). Если рассчитанное отношение шансов близко к единице, это означает, что связи нет и источники независимы (оценки по Линкольну-Петерсену достоверны).

В том случае, если найдены зависимые источники, их следует объединить друг с другом и рассматривать как один список. В дальнейшем анализ сводится к оценке численности популяции по методу Линкольна-Петерсена с использованием объединенного списка и списка, имевшего наименьшую корреляцию с другими.

Проиллюстрируем эту методику на примере. В г. Москве было проведено исследование, направленное на оценку количества потребителей инъекционных наркотиков (детальное описание методологии см. [18]). Были взяты данные из трех источников — результатов экспертизы наркологического опьянения, списка лиц, состоящих на учете в наркологическом диспансере и данных вызовов скорой помощи в связи с передозировкой наркотиков. Количество лиц, оказавшихся в разных списках, представлено на рис. 22.



Рис. 22. Результаты исследования по определению численности ПИН в г. Москве (Платт и соавт., 2005, не опубликовано)

На рисунке видно, что не было людей, которые бы присутствовали во всех трех списках, поэтому для расчета отношения шансов надо было воспользоваться откорректированной формулой, в которой к каждому значению численности прибавляют 0,5. На основании представленных данных отношение шансов для списков скорой помощи и наркодиспансера (по данным группы, которая была в списке экспертизы опьянения) составило

$$OR = \frac{0,5 * 2772,5}{35,5 * 89,5} = 0,44$$

Отношение шансов для списков скорой помощи и экспертизы опьянения составило

$$OR = \frac{0,5 * 1382,5}{89,5 * 1,5} = 5,14$$

Отношение шансов для списков наркодиспансера и экспертизы опьянения составило

$$OR = \frac{0,5 * 190,5}{35,5 * 1,5} = 1,78$$

Очевидно, что наибольшая корреляция наблюдается между списками скорой помощи и экспертизы опьянения, поэтому эти два списка следует объединить. Тогда полученные результаты сводятся к анализу двух списков, в одном из которых 3086 человек, а во втором — 1472 человека, и в обоих списках присутствуют 90 человек. По методу Линкольна-Петерсена численность ПИН тогда составляет

$$N = \frac{3086 * 1472}{90} = 50\,473$$

человека, а дисперсия

$$Var_N = \frac{3086 * 1472 * (3086 - 90) * (1472 - 90)}{90^3} = 2,58 * 10^7$$

95% доверительный интервал равен 40 600–60 400 человек. Соответственно, численность популяции ПИН в г. Москве, оцененная по методу Виттес-Линкольна-Петерсена, составляет от 41 до 60 тысяч человек с наиболее вероятной численностью около 50 тысяч⁴⁹.

Метод Виттес является достаточно простым и легко реализуемым даже в отсутствие значительных вычислительных мощностей и специалистов, обученных работать с системами статистической обработки данных. С другой стороны, он не устраняет полностью влияния существующей между источниками корреляции, поскольку может устранить лишь самый значительный из них (из трех списков делаются два, соответственно, возможная корреляция между остающимися источниками не устраняется). Желательно было бы иметь возможность построить математическую модель, которая бы позволяла учесть корреляции между всеми списками.

Такой моделью является логлинейная модель. В рамках этой модели логарифмы наблюдаемых частот моделируются как имеющие линейную связь друг с другом. Подобные модели широко распространены в статистике, и с их помощью моделируют различные процессы, такие, например, как смертность и заболеваемость (влияние факторов риска на эти процессы). Однако в ме-

⁴⁹ Более точная оценка по методу логлинейного моделирования дает значение от 35 до 52 тыс. с наиболее вероятным значением около 42 тыс. Хотя оценка по Виттес выше, говорить о достоверных различиях нельзя, поскольку доверительные интервалы перекрываются.

годах двойного охвата⁵⁰ аналитика интересует не влияние наблюдаемых факторов на частоты (наблюдаемые факторы в данном случае — списки), а предсказание количества лиц, оказавшихся вне списка. Для примера вернемся к табл. 7. Видно, что правая нижняя ячейка таблицы пуста, поскольку исследователям неизвестно, сколько ПИН не оказались ни в том, ни в другом списке. При обычном логлинейном моделировании были бы известны все четыре частоты и изучалась связь между факторами, образующими таблицу. В методах же двойного охвата будет строиться модель, описывающая данные таблицы, и на основании этой модели будут стараться предсказать содержимое отсутствующей ячейки. Попробуем построить логлинейную модель на примере табл. 7. Для простоты обозначим, что, если индивид присутствует в списке, то значение переменной, обозначающей этот список, составляет 1, а если отсутствует — то 0. Обозначим первый список А, а второй список В. Обозначим также частоты в таблицах латинскими буквами так, что частота в левой верхней ячейке (ранее обозначавшаяся как x_{12}) будет обозначена буквой а, частота для случая, когда лица попали во второй список, но отсутствуют в первом — b, случай, когда лица отсутствовали во втором списке, но присутствовали в первом — с. Количество тех, кто отсутствует в обоих списках, обозначим буквой d. Из определения логлинейной модели следует, что

$$\ln(a) = \alpha * A + \beta * B + \text{const} = \alpha * 1 + \beta * 1 + \text{const} = \alpha + \beta + \text{const}$$

$$\ln(b) = \alpha * A + \beta * B + \text{const} = \alpha * 0 + \beta * 1 + \text{const} = \beta + \text{const}$$

$$\ln(c) = \alpha * A + \beta * B + \text{const} = \alpha * 1 + \beta * 0 + \text{const} = \alpha + \text{const}$$

$$\ln(d) = \alpha * A + \beta * B + \text{const} = \alpha * 0 + \beta * 0 + \text{const} = \text{const}$$

На основании первых трех уравнений можно оценить три неизвестных параметра (коэффициенты α и β для списков и постоянный член уравнения const). На самом деле оценка коэффициентов для списков нас не очень интересует, а вот то, что нам нужно — это постоянный член уравнения. Он равен логарифму количества лиц, не попавших ни в один из списков и, соответственно, оценив его, можно оценить общую численность популяции $N = a + b + c + d$.

После этого теоретического введения можно попробовать решить систему уравнений для данных из табл. 7. Решение этой системы уравнений⁵¹ показывает, что $\text{const} = 10,34711$. Соответственно, взяв антилогарифм от этой величины (возводя число e в степень), получаем, что ни в один из списков не попало 31 167 человек. Всего же в популяции Бангкока тогда имеется $N = 171 + 1369 + 3893 + 31\ 167 = 36\ 600$ ПИН (число точно совпадает с расчетами по методу Линкольна-Петерсена).

Конечно, этот пример был очень простым, но он поясняет основную идею метода логлинейного моделирования. На основании наблюдаемых частот мы создаем систему уравнений, описывающих распределение наблюдаемых логарифмов частот, а затем пытаемся определить, чему равняется ненаблюдаемая частота (это всегда будет постоянный член уравнения логлинейной модели). Чрезвычайно важно то, что в модели можно предусмотреть взаимодействия между данными в разных списках и таким образом смоделировать связи, существующие между ними.

Надо, однако, отметить, что в приведенном выше примере для двух списков мы немного покривили душой. Дело в том, что реально полная логлинейная модель должна была бы включать в себя еще возможность взаимодействия между факторами А и В (т.е. учитывать возможность того факта, что между вероятностями попадания в списки А и В существует корреляция). Если бы это было учтено, то полная модель для двух списков стала бы такой:

⁵⁰ вернее, множественного охвата

⁵¹ например, с помощью процедуры REG системы SAS

$$\ln(a) = \alpha * A + \beta * B + \gamma * A * B + \text{const}$$

Очевидно, что в этом случае при не полностью заполненной таблице мы бы получили три уравнения с четырьмя неизвестными и не смогли бы их решить. Нас спасло предположение о независимости списков А и В. Иными словами, при использовании методов двойного охвата невозможно оценить полную логлинейную модель и всегда предполагается, что взаимодействие самого высокого уровня отсутствует. С другой стороны, если в распоряжении аналитика есть три списка, то он располагает семью значениями частот и, соответственно, может рассчитывать на построение семи уравнений с семью неизвестными. Если обозначить списки А, В и С, то, соответственно, можно оценить влияние каждого списка по отдельности (А, В и С) и их попарные взаимодействия (А * В, А * С, В * С). Оценить влияние всех трех факторов вместе не представляется возможным, поэтому предполагаем, что все три фактора одновременно друг с другом не коррелируют. Понятно, что это предположение не основано ни на чем, кроме нашего желания определить неизвестную численность популяции, и единственной гарантией его справедливости будет попытка исследователя найти списки, которые с минимальной вероятностью коррелируют друг с другом (т.е. не брать очевидно связанные списки). Если у аналитика в распоряжении есть большое количество списков, то необходимо выбрасывать только тот показатель в модели, который предполагает зависимость их всех друг от друга, а вероятность этого с ростом числа списков снижается. Поэтому, чем больше списков есть у аналитика, тем точнее будет оценка ненаблюдаемой популяции.

Возможность построения модели с большим количеством параметров (например, семь параметров для трех списков) базируется еще на одном предположении — что все показатели измерены абсолютно точно, т.е. если мы повторим эксперимент с получением списков из тех же источников, мы получим абсолютно такие же цифры (вернее, соотношение между значениями будет одинаковым). Это, естественно, является весьма сомнительным допущением. Однако для модели

$$\ln(a) = \alpha * A + \beta * B + \gamma * A * B + \text{const} + \varepsilon,$$

где ε — показатель, отражающий колебания значений в результате ошибки, у нас явно не хватает данных. В случае с двумя списками нам делать нечего, однако, если у нас есть три или более списков, то попытаться оценить ошибку можно. Для этого следует вначале установить, так ли уж необходимо иметь модель со всеми возможными взаимодействиями (конечно, без взаимодействия всех факторов вместе это оценить нельзя). Для этого надо последовательно построить ряд моделей — от самых простых к самым сложным, например, для трех списков простейшая модель будет выглядеть так:

$$\alpha * A + \beta * B + \gamma * C + \text{const} + \varepsilon$$

Для простоты записи в дальнейшем мы не будем писать α или β , а просто будем указывать наименование списка, предполагая, что для каждого подобного показателя мы оцениваем множитель. Соответственно, формула выше для простейшей модели в случае трех списков будет выглядеть так:

- А В С

Более сложные модели тогда определяются таким образом:

- А В С АВ
- А В С АС
- А В С ВС

- A B C A B A C
- A B C A B B C
- A B C B C A C
- A B C A B B C A C

Задумаемся на секунду, что происходит, когда от более простой модели мы переходим к более сложной. Фактически устраняется один из источников неопределенности (тот факт, что не учитывалось, например, взаимодействие A и B). Где этот источник неопределенности находился в более простой модели? В показателе ошибки ε . Соответственно, если мы оценим, насколько увеличился показатель ошибки при переходе к более простой модели, мы будем в состоянии оценить, насколько важным был источник неопределенности, внесенный в модель в явном виде. Например, если после прекращения учета взаимодействия A и B уровень неопределенности увеличился лишь незначительно, то очевидно, что взаимодействие AB не несет значимой информации в данной модели и его можно не учитывать. Таким образом, сравнивая более простые модели с более сложными, можно установить, какие взаимодействия являются значимыми, а какие — нет и, соответственно, подобрать наиболее простую и вместе с тем наиболее информативную модель.

Таблица 8

Распределение числа ПИН г. Тольятти по источникам информации

Списки задержанных	Списки наркодиспансера			
	есть		нет	
	списки СПИД-центра		списки СПИД-центра	
	есть	нет	есть	нет
есть	80	93	37	906
нет	894	1861	1351	

Для иллюстрации воспользуемся данными по определению численности ПИН в одном российском городе, которое проводилось с использованием трех списков — данных наркодиспансера, списков задержанных милицией наркоманов и данных СПИД-центра [18]. Всего в трех списках было 5222 человека, распределение по спискам представлено в табл. 8.

Эти данные можно записать в виде такой программы для ввода в систему SAS:

1. DATA tol;
2. INPUT narko aids police count;
3. CARDS;
4. 1 1 1 80
5. 1 1 0 894
6. 1 0 1 93
7. 1 0 0 1861
8. 0 1 1 37
9. 0 1 0 1351
10. 0 0 1 906
11. 0 0 0 .
12. ;
13. RUN;

Обратите внимание, что в последней строке ввода данных указано, что у нас нет данных по количеству лиц, оказавшихся вне всех трех списков (это будет необходимо для расчета этого числа). Теперь можно начинать логлинейное моделирование при помощи процедуры GENMOD. При выборе модели придется проанализировать все перечисленные выше модели и оценить их значимость с точки зрения более простых моделей. Для этого нам надо будет найти в распечатке, описывающей полученную модель, раздел, показывающий остаточную вариативность (логарифм правдоподобия, LL), и выписать его значение. Простейшая модель будет запрограммирована так:

1. PROC GENMOD;
2. MODEL count= narco aids police/
3. DIST=POISSON PREDICTED LRCI CL;
4. RUN;

Опции PREDICTED и CL позволяют оценить численность и доверительные интервалы численности ненаблюдаемой популяции по каждой из моделей (эти опции просят систему рассчитать предсказанные значения численностей для всех ячеек, в том числе той, где данные отсутствуют — вот зачем нужно было вводить пустое значение в файле данных). Выпишем результаты проведения анализа в табл. 9.

Для того, чтобы определить качество модели, рассчитывают разность логарифмов правдоподобия и делят полученную величину на разность степеней свободы (df), показывающих, какие еще варианты модели могли бы быть использованы (последняя модель является полной и там уже степеней свободы нет) [4]. Однако эта методика недостаточно хорошо работает, если количество наблюдений относительно велико, поскольку любые, даже незначительные изменения в качестве предсказания окажутся достоверными. Поэтому для отбора модели был предложен ряд показателей, из которых одним из наиболее адекватных является Байесовский информационный критерий⁵² (BIC). Формула для оценки модели выглядит так:

$$BIC = 2 * (LL_p - LL_r) - df_r * \ln(N),$$

где LL_p — логарифм правдоподобия для полной модели и LL_r — логарифм правдоподобия для упрощенной модели, df_r — количество степеней свободы для упрощенной модели, а N — общее количество наблюдений.

Таблица 9

Анализ логлинейных моделей

Модель	LL	df	Численность
A B C	31 392	3	4965
A B C AB	31 660	2	17 714
A B C AC	31 467	2	3821
A B C BC	31 469	2	4001
A B C AB BC	31 660	1	18 130
A B C AB AC	31 671	1	33 081
A B C BC AC	31 579	1	2812
A B C AB BC AC	31 678	0	59 237

⁵² Другие показатели для отбора модели включают, в частности, Информационный критерий Акаике (AIC), который рассчитывается по следующей формуле: $2 * (LL_p - LL_r) - 2 * df_r$.

Сделаем эти расчеты и соберем их в таблицу (см. табл. 10).

Таблица 10

Оценка логлинейных моделей

Модель	ВІС
А В С АВ ВС	6,4
А В С АВ АС	190,4
А В С ВС АС	28,4
А В С АВ	402,8
А В С АС	406,8
А В С ВС	20,8
А В С	549,2

Считается, что наиболее адекватная модель имеет самый низкий ВІС. В данном случае — это полная модель (А В С АВ ВС АС), однако еще две модели — А В С АВ ВС и А В С ВС также имеют невысокие значения ВІС. Надо отметить, что информация о разных моделях — это все, чем может помочь компьютерная техника аналитику. Окончательный отбор производится не автоматически, а путем анализа реальности получаемых результатов, сложности модели и других параметров. Так, в приведенном выше примере наиболее адекватная со статистической точки зрения модель — полная модель. Однако данная модель предсказывает наличие в данном городе почти 60 тысяч ПИН, не попавших ни в одни списки. Это означает, что чувствительность каждого списка (определяемая как отношение количества лиц в списке к оценке общей численности популяции) крайне мала. Поэтому наиболее адекватным будет выбрать другую модель⁵³ — А В С АВ ВС — с немного более высоким значением ВІС, но зато меньшим количеством параметров. Согласно этой модели численность не попавших в списки ПИН составляет 18 тыс. человек (95% ДИ = = 14–22 тыс. человек). Иногда имеет смысл попытаться разделить наблюдения, например по полу и возрасту и оценить численность этих популяционных групп по отдельности, а затем посмотреть, совпадает ли суммарная численность популяции в результате использования ряда отдельных моделей с оценками по одной суммарной модели. Альтернативным вариантом является построение модели с использованием дополнительной информации в виде включаемых в модель переменных, таких, например, как пол, возраст, место проживания и т.п. В любом случае надо понимать, что оценка по методу двойного охвата при отсутствии уверенности в независимости списков (по моделям) является лишь ориентировочной, однако полученные ориентировочные значения можно использовать в моделировании с использованием различных сценариев — от наиболее благоприятного (например, с минимальным количеством ПИН), до наиболее неблагоприятного (с максимальной оценкой ПИН). Самая большая ошибка, которую может сделать аналитик — это считать, что оценка численности ненаблюдаемой популяции является точной до последнего человека, даже если она получена при помощи сложной математической модели⁵⁴.

Приведенные выше данные показывают, что наличие большого количества списков, желательно не зависящих или мало зависящих друг от друга, позволяет повысить точность оценки численности популяции путем моделирования взаимоотношений между списками.

⁵³ Как было показано, полные модели часто завышают размер популяции [8].

⁵⁴ Здесь вспоминается высказывание У. Черчилля о демократии, что это плохая форма правления, однако альтернативы — еще хуже. В нашем случае альтернативой является полное отсутствие оценок численности популяции и невозможность планировать в принципе.

Иногда возникает ситуация, при которой нет разных списков, а имеется один список, в который, однако, люди могут попадать один, два или более раз. При наличии подобного списка можно предположить, что распределение количества пациентов подчиняется распределению Пуассона. Тогда задачей является определить, какое количество людей не попали в этот список ни разу. Подобный тип распределения называется распределением, обрезанным в нуле (поскольку мы предполагаем, что имеются лица, которые отсутствовали в списке вообще, но сколько их — мы не знаем). Для оценки численности популяции можно использовать следующее приближение:

$$N = \sum f_i + \frac{f_1^2}{2 * f_2},$$

где f_1 — количество лиц, обнаруженных в списке один раз, а f_2 — количество лиц, обнаруженных в списке два раза. В качестве примера приведем методику оценки количества ПИН на основании данных программы снижения вреда [11]. Количество лиц, которые посещали когда-либо центр обмена шприцев, составило 765 человек. Из них 473 посетили центр обмена шприцев только один раз и 97 человек — два раза. Используя описанную выше формулу, можно получить, что количество ПИН в городе составляет

$$N = 765 + \frac{473^2}{2 * 97} = 1918$$

Методика является достаточно простой, и ее можно использовать для экспресс-оценки в том случае, если данные с повторами легко доступны. Однако не стоит забывать, что эта методика будет давать большую ошибку, если вероятность прихода после первого визита увеличивается (человеку «нравится» посещать то место, откуда получены данные для формирования списка) или, наоборот, снижается (например, такое будет происходить в случае списков задержаний — часть будет осуждаться и явно не сможет попасть в список повторно). В первом случае оценки будут заниженными, а во втором — завышенными. Кроме того, если материал собирается на протяжении длительного промежутка времени, в случае открытой популяции ошибка, связанная с возможностью иммиграции и эмиграции, может быть значительной.

4.3. Методы для открытой популяции

Анализ открытой популяции осложняется тем, что не совсем понятно, какая именно численность популяции интересует аналитика. Понятно, что в открытой популяции (если она не является стационарной) численность меняется и, как было показано выше, когда-то стандартный метод Линкольна-Петерсена дает оценку численности популяции в начальный момент времени, а когда-то — в конечный. Поэтому методики оценки численности открытой популяции должны включать оценку скорости притока новых членов в популяцию и скорости оттока.

Таблица 11

Организация данных для оценки популяции по методу Бейли (тройного охвата)

Период	Время	Выявлено	«Помечено» новых	«Помеченных» из выявленных
0	0		s_0	
1	t_1	n_1	s_1	m_{01}
2	t_1+t_2	n_2		m_{012}, m_{02}, m_{12}

Для ответа на вопрос, не является ли популяция открытой, было разработано достаточно много методов, и мы начнем с простейшего из них — метода Бейли, предложенного еще в 1951 году

(другое название метода — метод тройного охвата (triple catch)). Этот метод удобен для пилотных исследований и экспресс-оценки [12]. Для его использования необходимо располагать данными по трем различным временным периодам. В каждый из этих периодов отмечают, сколько человек всего было проанализировано в этот период (n) и сколько из них были выявлены в предшествующие периоды (m). Полученные результаты можно свести в следующую таблицу (табл. 11).

Очевидно, что лица, помеченные в нулевой период времени, могли быть обнаружены в выборках первого и второго периодов. Для первого периода количество лиц с метками из нулевого периода равно m_{01} . Для второго периода количество лиц с метками из нулевого периода равно $r_{02} = m_{02} + m_{012}$, а из первого — $r_{12} = m_{12} + m_{012}$. Понятно, что количество вновь помеченных в первый период равно $s_1 = n_1 - m_{01}$. Располагая этой информацией, можно оценить численность популяции в первом периоде:

$$N = \frac{n_1 * s_1 * r_{02}}{m_{01} * r_{12}} = \frac{(n_1 - m_{01}) * n_1 * (m_{02} + m_{012})}{m_{01} * (m_{12} + m_{012})}$$

Дисперсия полученной оценки численности популяции составляет

$$Var_N = N^2 * \left(\frac{1}{r_{12}} + \frac{1}{r_{02}} + \frac{1}{m_{01}} + \frac{1}{n_2} \right)$$

Полученная оценка дисперсии может быть использована для расчетов 95% доверительного интервала, как описано выше для оценок по Линкольну-Петерсену. Однако следует помнить, что приведенные формулы справедливы только для относительно больших выборок. В тех случаях, когда выборки невелики, можно использовать формулы с так называемым фактором коррекции Бейли:

$$N = \frac{(n_1 + 1) * s_1 * r_{02}}{m_{01} * r_{12}}$$

В этом случае выражение для дисперсии меняется следующим образом:

$$Var_N = N^2 - \frac{(s_1)^2 * (n_1 + 1) * (n_1 + 2) * (r_{02} - 1) * r_{02}}{(m_{01} + 1) * (m_{01} + 2) * (r_{12} + 1) * (r_{12} + 1)}$$

Наиболее интересным для исследователя является не численность популяции, а возможность оценить скорость изменения численности популяции. Скорость снижения численности популяции, которая включает в себя все выбытия (смертность и эмиграцию), рассчитывается по следующей формуле:

$$\gamma = \frac{1}{t_1} \ln \left(\frac{s_1 * r_{02}}{s_0 * r_{12}} \right)$$

Скорость роста популяции (появление новых членов, «рождаемость» и иммиграция) оценивается по формуле

$$\beta = \frac{1}{t_2} \ln \left(\frac{m_{01} * n_2}{n_1 * r_{02}} \right)$$

Зная численность популяции в первый период времени и скорость прироста и выбытия членов популяции, можно рассчитать численность популяций в начальный момент времени и в нулевой момент (и, теоретически, в любой другой момент времени, если забыть об опасностях экстраполяции).

Проиллюстрируем данный метод на гипотетическом примере. В результате скрининга в городе N было выявлено 1200 ВИЧ-инфицированных пациентов, которые были поставлены на диспансерный учет. Через год было дополнительно выявлено еще 500 случаев ВИЧ-инфекции, однако из тех, кто был выявлен в первый год, контакт сохранился лишь с 840 пациентами. Соответственно, на втором году на учете состояли $500 + 840 = 1340$ человек. На третий год на диспансерном учете продолжали состоять (были в контакте) 700 человек из выявленных в ходе первоначального скрининга и 400 из тех, кто был выявлен на втором году. Кроме того, на третьем году было выявлено 300 новых случаев ВИЧ-инфекции, соответственно, на учете состояли $700 + 400 + 300 = 1400$ человек. Для удобства расчетов эти цифры можно представить так (табл. 12).

Таблица 12

Исходные данные для расчета по методу тройного охвата

Период	0	1	2
0	1200		
1	840	1340	
2	700	400	1400

В диагональных элементах таблицы находится численность ВИЧ-инфицированных, которые состояли на диспансерном учете (с ними в текущем году имелся контакт). Используя приведенные выше формулы, численность ВИЧ-инфицированных на втором году составляет

$$N = \frac{(1340 - 840) * 1340 * 700}{840 * 400} = 1396$$

Расчетная величина дисперсии для случая значительных размеров популяции составляет 11 428, а 95% доверительный интервал — от 1186 до 1605 человек. Показатель скорости роста количества инфицированных составляет

$$\beta = \ln\left(\frac{840 - 1400}{1340 * 700}\right) = 0,226$$

Следует обратить внимание, что, поскольку анализировались данные за один год, то оценка β не включала в себя деления на какое-либо число (надо было делить на единицу), однако можно было, например, перевести скорость в месячные показатели путем деления на 12. Для того, чтобы перевести показатель скорости роста числа инфицированных в более удобный, необходимо взять антилогарифм от произведения коэффициента β и времени, для которого рассчитывается скорость. Например, за один год прирост количества случаев ВИЧ-инфекции будет составлять $e^\beta = 1,25$. Иными словами, за год количество случаев ВИЧ-инфекции в среднем растет на 25%. Оценка скорости выбытия (смерть или потеря контакта вследствие недоступности для системы здравоохранения) проводится путем расчета коэффициента γ :

$$\gamma = \ln\left(\frac{(1340 - 840) * 700}{1200 * 400}\right) = -0,316$$

Так же, как и в случае скорости прироста, вероятность остаться в наблюдаемой группе в течение года рассчитывается путем взятия антилогарифма $e^\gamma = 0,73$. Таким образом, ежегодно из контакта с системой здравоохранения выходили 27% всех ВИЧ-инфицированных и 73% оставались в контакте. Зная эти показатели, можно оценить численность ВИЧ-инфицированных в первый и третий года. Обозначим за N_0 численность ВИЧ-инфицированных в первый год. Поскольку в контакте остается 73% всех ВИЧ-инфицированных, значит, ко второму году под наблюде-

нием должны были остаться $N_0 * 0,73$ пациентов. Кроме того, средний прирост количества инфицированных составляет 25% в год, соответственно, количество новых случаев определяется соотношением $N_0 * 0,25$. Отсюда следует, что количество ВИЧ-инфицированных ко второму году составит $N = N_0 * (0,73 + 0,25)$ человек. Из приведенных выше расчетов известно, что численность популяции ВИЧ-инфицированных на втором году составляет 1396 человек. Соответственно, численность ВИЧ-инфицированных в первый год должна быть $N_0 = N / 0,98 = 1424$ человека. Аналогичным образом можно оценить и численность популяции в третий год. Надо заметить, что метод тройного охвата базируется на предположении о том, что скорости выбытия и вхождения в популяцию остаются неизменными, что может достаточно часто нарушаться (например, в описанном выше примере репродуктивное число для пациентов первого года составило 0,41, а для пациентов второго — 0,22). В таких случаях оценки скорости изменения популяции являются сомнительными и следует пользоваться другими методами, которые полагаются на значительно большее количество временных точек.

Наиболее часто используемым для полного анализа открытых популяций является метод Джоли-Себера (Jolly-Seber), который требует от исследователя детальной информации о том, в какой период субъект первый раз попал в выборку и как часто и в каких последующих выборках он появлялся. Иными словами, деперсонализованные данные, часто используемые в других моделях двойного охвата, в данном случае являются неприемлемой. Вместе с тем, располагая персонифицированными данными и применяя метод Джоли-Себера, исследователь получает возможность оценить динамику популяции, рождаемость (и иммиграцию) и смертность (плюс эмиграцию). Для этого метода следует располагать как минимум тремя временными точками, однако лучше, чтобы временных точек, когда проводилась оценка, было бы больше. Для описания этого метода воспользуемся гипотетической ситуацией⁵⁵.

Недавно начавшая работать программа снижения вреда хочет оценить размер популяции ПИН, пользующейся ее услугами, и динамику численности этой популяции. Для этого регистрируются клиенты программы и затем еженедельно подводятся итоги. Каждую неделю определяется, сколько всего клиентов посетили программу, какое количество из них были новыми, а какое — старыми. При этом также определяется, когда впервые человек пришел в программу и появлялся ли он в программе в предшествующие недели. Этой информации, в принципе, достаточно для построения модели открытой популяции по Джоли-Себеру⁵⁶.

Для построения модели надо вначале организовать все данные в таблицы по периодам первого контакта пациентов с центром. Например, на второй неделе с центром проконтактировало 12 новых клиентов (кроме того, в центр пришло 12 клиентов, которые впервые пришли в центр на первой неделе, но пока, разбирая первую стадию формирования суммарных таблиц, эту информацию мы проигнорируем). Рассматривая 12 клиентов, которые впервые обратились в центр на второй неделе, выясняется, что четверо из них затем появились в центре на третьей неделе, пятеро — на четвертой и 7 человек на пятой неделе. Данные по визитам можно свести в следующую таблицу (табл. 13).

Из таблицы 13 видно, что клиенты под номерами 29 и 39 посещали центр каждую неделю, клиент под номером 30 не приходил в центр на четвертой неделе, а клиент под номером 38, наоборот, пропустил только третью неделю. Для дальнейших расчетов необходимо оценить количество людей, которые посетили центр в период, непосредственно следующий за первым контактом с центром (т.е. в данном случае на третьей неделе⁵⁷ — из таблицы видно, что таких было четыре человека). Далее анализируется информация только по тем клиентам, которые

⁵⁵ В данном случае сознательно приводятся разные примеры по сравнению с методикой тройного охвата, чтобы продемонстрировать диапазон применения методов анализа численности ненаблюдаемых открытых популяций.

⁵⁶ Для облегчения освоения этого метода исходные таблицы, на основании которых производятся расчеты для этого примера, приведены в Приложении.

⁵⁷ Второй период после первоначального контакта на второй неделе.

не проконтактировали с центром во втором периоде, но проконтактировали в третьем (таких на четвертой неделе было трое). Два человека не посещали центр на третьей и четвертой неделе, но появились на пятой. Итак, общее число лиц, которые контактировали с центром после первоначального контакта, составило $4 + 3 + 2 = 9$ человек, и трое более вообще не появились в центре.

Таблица 13

Данные по 12 клиентам центра для модели Джоли-Себера

IDN	Неделя				
	Первая	Вторая	Третья	Четвертая	Пятая
29	-	1	1	1	1
30	-	1	1	-	1
31	-	1	1	-	-
32	-	1	-	-	-
33	-	1	-	-	-
34	-	1	-	-	1
35	-	1	-	-	-
36	-	1	-	1	1
37	-	1	-	1	-
38	-	1	-	1	1
39	-	1	1	1	1
40	-	1	-	-	1

Оценка численности и динамики популяции по методу Джоли-Себера облегчается, если вначале свести результаты наблюдения в суммарную таблицу, аналогичную той, что показана в табл. 14. В верхней части таблицы указаны суммарные результаты по каждому периоду, в котором проводится анализ (в нашем примере — недели). Вначале указывается общее количество лиц, которое обратилось в центр на этой неделе (n_i). Следующая строка показывает число тех, кто уже являлся клиентами центра в предшествующие периоды (m_i) и количество новых лиц, обратившихся в центр (u_i)⁵⁸. Понятно, что общая численность обратившихся в центр равна сумме новых и «старых» клиентов. Далее идет количество лиц, которые покинули центр (s_i) в том смысле, что они вернулись назад, а не были удалены из популяции, например, потому, что были срочно госпитализированы. В большинстве случаев общая численность пришедших и центр и покинувших его клиентов будет равна, однако модель не требует подобного допущения.

Далее идет самая сложная для заполнения часть таблицы, которая требует знания информации о том, когда каждый клиент последний раз был в центре перед настоящим приходом. Очевидно, что для первой недели эти данные не имеют смысла (все пришли первый раз), однако уже для второй можно проводить расчеты. В строке указан период, в который клиент последний раз обращался в центр, а в столбце — тот период, который анализируется. Так, например, если анализируется второй период, то из 24 клиентов, обратившихся в центр на второй неделе, 12 были повторными, которые уже посещали центр на первой неделе, соответственно, число 12 заносится на пересечение первой строки (когда был последний раз) и второго столбца (анализируемый период). Поскольку более ранних периодов не было, нижняя часть этого столбца остается не заполненной. В третий период (на третьей неделе) в центр также обратились 24 клиента, и из них 2 последний раз обращались в центр на первой неделе. Соответственно, число 2 заносится на пересечении первой строки и второго столбца. При изучении количества лиц, которые проконтактировали с центром на второй и третьей неделе, следует учитывать тех, кто впервые обратились в центр на первой и второй неделях. Таких людей было семеро из числа впервые

⁵⁸ На пятой неделе количество новых клиентов и их судьба уже не анализируются, анализируется только информация по клиентам, впервые обратившимся в первые четыре недели.

посетивших центр на первой неделе и четверо из числа впервые обратившихся на второй (см. также табл. 13). Сумма этих значений и заносится на пересечении второй строки и третьего столбца.

Таблица 14

Суммарная таблица результатов оценки визитов для метода Джоли-Себера

Период последнего визита	Неделя визита				
	1	2	3	4	5
1		12	2	2	1
2			11	4	4
3				11	3
4					15
n_i	28	24	24	22	28
m_i	0	12	13	17	22
u_i	28	12	11	5	
s_i	28	24	24	22	

Аналогичным образом заполняется вся таблица, которая теперь может служить для расчета двух важнейших вспомогательных величин — числа лиц, обратившихся на i -ой неделе и еще хоть раз посетивших центр потом (r_i) и количества лиц, которые посетили центр после первого визита (z_i).

Количество лиц, которые посетили центр впервые на i -ой неделе и затем посетили центр хотя бы раз, равно сумме значений в строке, соответствующей этой неделе. Показатель z_i вычисляется как сумма всех значений ячеек, расположенных в строках выше i -ой и столбцах, расположенных правее i -ого. Так, для первой недели показатель z_1 не определен, поскольку строк выше нет. Для второй недели его значение равно $2 + 2 + 1 = 5$. Для третьей недели уже $2 + 1 + 4 + 4 = 11$. И, наконец, для четвертой недели оно равно сумме ячеек пятого столбца с первой по третью строки, т.е. $1 + 4 + 3 = 8$.

После заполнения таблицы расчеты численности популяции для каждого периода являются достаточно простыми. Вначале оценивается вспомогательная величина M_i :

$$M_i = \frac{(s_i + 1) * z_i}{(r_i + 1)} + m_i$$

Далее численность популяции рассчитывается по следующей формуле:

$$N_i = \frac{(n_i + 1) * M_i}{(m_i + 1)}$$

В анализируемом примере численность популяции, из которой приходили клиенты программы, составляла на первой неделе 35 человек, на второй неделе 56 человек и на третьей неделе 36 человек.

Помимо оценки численности популяции, модель Джоли-Себера, как и другие модели открытых популяций, позволяет установить динамику в этой популяции — частоту иммиграции

и эмиграции. Эмиграция оценивается через количество лиц, остающихся в популяции (иными словами, дополнение до 1 (100%) от количества оставшихся). Количество же остающихся в популяции измеряется как процент лиц, находившихся в популяции в предшествующий период, которые находятся в ней в нынешний период:

$$\Phi_i = \frac{M_{i+1}}{M_i + s_i - m_i}$$

Для первого периода времени ($i = 1$), M_1 устанавливается равной нулю. Тогда в описываемом примере $\Phi_1 = 0,652$, $\Phi_2 = 1,036$ (значения Φ могут немного превышать единицу из-за ошибок измерения, как в приводимом примере, однако значительное превышение единицы чаще всего указывает на ошибку данных) и $\Phi_3 = 0,673$. Эти результаты можно трактовать как указание на то, что каждую неделю центр «терял» около 40% своих клиентов.

Оценка иммиграции проводится путем подсчета количества новых лиц, на которых увеличился размер популяции. Расчет производится по формуле:

$$B_i = N_{i+1} - \Phi_i * (N_i - n_i + s_i).$$

Количество новых лиц в популяции определяется как разность между численностью популяции в последующий период и численностью популяции в настоящий период, помноженной на частоту удержания лиц в программе. Кроме того, из численности популяции вычитается количество лиц, которые покинули популяцию в результате контакта с программой (разность n_i и s_i). Если количество лиц, пришедших с визитом в программу, равно количеству лиц ушедших, то разность $n_i - s_i$ обращается в ноль. Используя эту формулу, можно выяснить, что между первой и второй неделями действия программы численность популяции увеличилась на 30 человек.

Конечно, проводить ручные расчеты, если количество анализируемых периодов достаточно велико, а также велико и количество субъектов, находящихся под наблюдением, достаточно сложно. Поэтому желательно использовать программное обеспечение, которое позволяет автоматизировать эту работу. За прошедшие годы было создано достаточно большое количество специализированных программ, таких как MARK или SURPH, обычно используемых специалистами-экологами для определения численности популяции. Однако мы проиллюстрируем автоматизацию расчетов с помощью системы SAS, на которой приводились другие примеры в данной работе. Для анализа потребуются написать программу, которая будет оперировать с массивом данных, и для этой цели наилучшим образом подходит встроенный матричный язык SAS, называемый IML (Interactive Matrix Language). Сама программа просто автоматически выполняет те действия, которые были описаны выше:

1. PROC IML;
2. USE caprec;
3. READ ALL INTO catch;
4. m=J(NCOL(catch),NCOL(catch),0);
5. u=J(NCOL(catch),1,0);
6. r=J(NCOL(catch),1,0);
7. marked=J(NCOL(catch),1,0);
8. n=J(NCOL(catch),1,0);
9. z=J(NCOL(catch),1,0);
10. mmm=J(NCOL(catch),1,0);

```

11. nn=J(NCOL(catch),1,0);
12. alfa=J(NCOL(catch),1,0);
13. Phi=J(NCOL(catch),1,0);
14. B=J(NCOL(catch),1,0);
15. DO i=1 TO NROW(catch);
16. k=0;
17.     DO j=1 TO NCOL(catch);
18.     IF catch[i,j]=1 THEN
19.     IF k=0 THEN DO;
20.     u[j]=u[j]+1;
21.     k=j;
22.     END;
23.     ELSE DO;
24.     m[k,j]=m[k,j]+1;
25.     k=j;
26.     marked[j]=marked[j]+1;
27.     END;
28.     END;
29. END;
30. DO j=1 TO NCOL(catch);
31.     r[j]=m[j,+];
32. END;
33. DO k=1 TO NCOL(catch);
34. DO j=1 TO NCOL(catch);
35.     DO i=1 TO NCOL(catch);
36.     IF j>(k) & i<(k) THEN z[k]=z[k]+m[i,j];
37.     END;
38. END;
39. END;
40. n=marked+u;
41. mmm=(n+1)#z/(r+1)+marked;
42. alfa=(marked+1)/(n+1);
43. nn=mmm/alfa;
44. DO i=1 TO NCOL(catch)-2;
45.     Phi[i]=mmm[i+1]/(mmm[i]+n[i]-marked[i]);
46.     B[i]=nn[i+1]-Phi[i]*nn[i];
47. END;
48. PRINT n marked u r z mmm nn Phi B;
49. QUIT;

```

Вначале данные вводятся в систему при помощи обычного шага данных DATA (не показано), а затем весь файл (саргес) передается в процедуру IML (команды USE и READ — первая сообщает IML, какой файл надо будет использовать, а вторая считывает все содержимое файла

в матрицу с именем catch. Далее определяются все необходимые матрицы данных, и с 15-й строки начинается расчет значений для массива, аналогичного табл. 14. Расчеты промежуточных показателей и численности популяции (nn) проводятся по приведенным выше формулам. Предпоследняя строка выводит на печать все суммарные данные, такие как количество лиц, посетивших центр в данную неделю (n), количество тех из них, кто ранее посещал центр (marked), количество новых клиентов⁵⁹ (u), а также вспомогательные величины r , z и M (обозначенная mmm). Заключают распечатку численность популяции в этом периоде, выживаемость и иммиграция (Φ (Phi) и B , соответственно).

Надо заметить, что в разбираемом примере как численность популяции, так и ее динамика относятся не ко всем ПИН, а лишь к «популяции» программы снижения вреда. Иными словами, если ранее в городе были получены оценки численности популяции другими методами (например, предполагая закрытость популяции по спискам скорой помощи, наркодиспансера и центра СПИД, как описано ранее), то, сравнивая два полученных значения, можно оценить степень охвата целевой популяции.

Более того, динамика популяции также может оказаться полезной для оценки качества деятельности программы. Так, если «эмиграция» является высокой, это означает, что клиенты не удовлетворены качеством и спектром услуг и поэтому мало задерживаются в программе. В том же случае, если при низком охвате наблюдается низкая «иммиграция», это показывает недостаточную осведомленность потенциальных клиентов о деятельности программы.

Приведенное обсуждение, очевидно, применимо не только к программам снижения вреда, но и к другим программам системы здравоохранения, включая, например, оценку динамики использования услуг здравоохранения в первичном звене. Кроме того, очевидно, что эти методы могут быть напрямую применены для оценки количества лиц, больных заразными заболеваниями, поскольку, в отличие от методик, ориентированных на анализ закрытой популяции, они легко учитывают прирост количества больных в результате распространения заразного заболевания и уменьшение пула больных в результате выздоровления.

Таким образом, методики определения численности открытой популяции, в особенности методы множественного охвата, к которым относится метод Джоли-Себера, являются важным инструментом в работе эпидемиолога и специалиста общественного здоровья.

⁵⁹ Обратите внимание, что в последнем временном интервале количество новых субъектов обычно не учитывается.

Заключение

Широкое распространение компьютерных технологий революционизировало многие области человеческой деятельности. Возможности компьютеров по решению задач путем перебора различных вариантов открыли новые направления в эпидемиологии. Если раньше решаемые задачи ограничивались лишь теми, для которых можно было найти аналитическое решение, то сейчас эти ограничения сняты и стохастические методы моделирования все больше и больше выдвигаются на первые планы в области изучения особенностей эпидемических кривых. Возможности по накоплению и анализу больших массивов информации также увеличились, и теперь модели могут строиться с учетом большого количества разнообразных параметров и изучать влияние их варьирования на течение эпидемического процесса.

Вместе с тем нельзя сказать, что математические методы в эпидемиологии занимают то место, которое могли бы с учетом их возможностей и той информации, которую они предоставляют. В качестве примера можно отметить, что очень сложно найти российские исследования, оценивающие эпидемические кривые для ВИЧ-инфекции методами обратного расчета или изучающие численность популяций риска методами двойного охвата⁶⁰. Возможно, что причиной тому является отсутствие подготовленных кадров, поскольку математическая эпидемиология практически не фигурирует ни в программах медицинских вузов (даже по специальности «медико-профилактическое дело»), ни в программах послевузовской подготовки. Учитывая тесную связь современной математической эпидемиологии с биостатистикой, можно отметить, что и этот предмет не является обязательным во всех медицинских вузах. Эта ситуация должна измениться, и тогда инструменты математической эпидемиологии придут на помощь эпидемиологам в их повседневной работе по изучению эпидемического процесса и профилактике заразных болезней человека.

⁶⁰ Хотя классический труд по математическому моделированию эпидемий Р. Андерсона и Р. Мэя был переведен на русский язык [1].

Список литературы

1. Андерсон, Р. Инфекционные болезни человека. Динамика и контроль / Р. Андерсон, Р. Мэй. — Москва: Издательство «Мир», 2004. — С. 1–783.
2. Баррет, К. Виртуальная атака биотеррористов / К. Баррет, С. Юбанк, Д. Смит // В мире науки. — 2005. — № 6. — С. 38–45.
3. Плавинский, С. Математическое моделирование распространения инфекций, передающихся половым путем. Значение для общественного здравоохранения / С. Плавинский // Российский семейный врач. — 2002. — № 1. — С. 16–22.
4. Плавинский, С. Биостатистика. Планирование, анализ и представление биомедицинских данных с помощью системы SAS / С. Плавинский. — Санкт-Петербург: Издательский дом МАПО, 2005. — С. 1–555.
5. Плавинский, С. Сексуальное поведение, венерические болезни и гетеросексуальная эпидемия ВИЧ-инфекции — некоторые результаты математического моделирования / С. Плавинский, А. Баринава, К. Разнатовский // Российский семейный врач. — 2007. — № 3. — С. 30–37.
6. Покровский, Б. Эпидемиология и профилактика ВИЧ-инфекции и СПИД / В. Покровский. — Москва: Медицина, 1996.
7. Федеральный научно-методический центр по профилактике и борьбе со СПИДом. ВИЧ-инфекция. Информационный бюллетень № 30 / Федеральный научно-методический центр по профилактике и борьбе со СПИДом. — Министерство здравоохранения и социального развития РФ, 2007.
8. Estimating infectious diseases incidence: validity of capture-recapture analysis and truncated models for incomplete count data / N. Van Hest, A. Grant, F. Smit et al. // *Epidemiol Infect.* — 2008. — Vol. 136, no. 1. — Pp. 14–22.
9. Ewald, P. W. *Evolution of Infectious Disease* / P. W. Ewald. — Oxford University Press, 1994.
10. Gisselquist, D. Efficiency of human immunodeficiency virus transmission through injections and other medical procedures: evidence, estimates, and unfinished business. / D. Gisselquist, G. Upham, J. J. Potterat // *Infection control and hospital epidemiology.* — 2006. — Vol. 27, no. 9. — Pp. 944–952.
11. Hay, G. Truncated Poisson analyses of data from the Riga needle exchange: Tech. rep. / G. Hay: University of Glasgow, 2003.
12. Henderson, P. *Practical Methods in Ecology* / P. Henderson. — Black-well Publishing, 2003.
13. Isham, V. *Celebrating Statistics: Papers in Honour of Sir David Cox on his 80th Birthday* / V. Isham / Ed. by A. Davison, Y. Dodge, N. Wer-muth. — Oxford University Press, 2005. — Pp. 5–36.
14. Meyer, W. *Concepts of Mathematical Modeling* / W. Meyer. — Mineola, New York: Dover Publications, Inc., 1984.

15. Molecular epidemiology: HIV-1 and HCV sequences from Libyan outbreak. / T. de Oliveira, O. G. Pybus, A. Rambaut et al. // *Nature*.— 2006. — Vol. 444, no. 7121. — Pp. 836–837.
16. Nosocomial outbreak of multiple bloodborne viral infections. / S. Yerly, R. Quadri, F. Negro et al. // *The Journal of infectious diseases*.— 2001. — Vol. 184, no. 3. — Pp. 369–372.
17. Nowak, M. A. *Virus Dynamics. Mathematical Principles of Immunology and Virology* / M.A. Nowak, R.M. May. — Oxford University Press, 2000.
18. The prevalence of injecting drug use in a russian city: implications for harm reduction and coverage / L. Piatt, M. Hickman, T. Rhodes et al. // *Addiction*. — 2004. — Vol. 99. — Pp. 1430–1438.
19. Rao, A. Incubation-time distribution in back-calculation applied to HIV/AIDS data in India / A. Rao, M. Kakehashi // *Mathematical Biosciences and Engineering*. — 2005. — apr. — Vol. 2, no. 2. — Pp. 263–277.
20. Risk of HIV-1 transmission for parenteral exposure and blood transfusion: a systematic review and meta-analysis. / R.F. Baggaley, M.C. Boily, R.G. White, M. Alary // *AIDS (London, England)*. — 2006. — Vol. 20, no. 6. — Pp. 805–812.
21. Wasserheit, J.N. The dynamic topology of sexually transmitted disease epidemics: implications for prevention strategies / J.N. Wasserheit, S.O. Aral // *J Infect Dis*. — 1996. — Oct. — Vol. 174 Suppl 2. — Pp. 201–213.
22. What percentage of the Cuban HIV-AIDS epidemic is known? / H. de Arazoza, R. Lounes, J. Perez, T. Hoang // *Rev Cubana Med Trop*. — 2003. — Vol. 55, no. 1. — Pp. 30–37.
23. Wittes, J. A generalization of the simple capture-recapture model with applications to epidemiological research. / J. Wittes, V. Sidel // *J Chronic Dis*. — 1968. — Vol. 21. — Pp. 287–301.

Приложение

Таблица визитов клиентов программы снижения по неделям (гипотетический пример)

Недели IDN	1	2	3	4	5
1	1	1	1	1	1
2	1	1	1	1	1
3	1	1	1	1	1
4	1	1	1	1	0
5	1	1	1	0	1
6	1	1	1	0	0
7	1	1	1	0	1
8	1	1	0	1	0
9	1	1	0	0	0
10	1	1	0	0	0
11	1	1	0	0	1
12	1	1	0	0	1
13	1	0	1	1	1
14	1	0	1	1	1
15	1	0	0	1	1
16	1	0	0	1	1
17	1	0	0	0	0
18	1	0	0	0	0
19	1	0	0	0	0
20	1	0	0	0	0
21	1	0	0	0	0
22	1	0	0	0	0
23	1	0	0	0	1
24	1	0	0	0	0
25	1	0	0	0	0
26	1	0	0	0	0
27	1	0	0	0	0
28	1	0	0	0	0
29	0	1	1	1	1
30	0	1	1	1	1
31	0	1	1	0	1
32	0	1	1	0	0
33	0	1	0	1	1
34	0	1	0	1	1
35	0	1	0	1	0
36	0	1	0	0	1
37	0	1	0	0	1
38	0	1	0	0	0
39	0	1	0	0	0
40	0	1	0	0	0
41	0	0	1	1	1
42	0	0	1	1	1
43	0	0	1	1	0
44	0	0	1	0	0
45	0	0	1	0	0
46	0	0	1	0	0
47	0	0	1	0	0

48	0	0	1	0	0
49	0	0	1	0	0
50	0	0	1	0	0
51	0	0	1	0	0
52	0	0	0	1	1
53	0	0	0	1	1
54	0	0	0	1	0
55	0	0	0	1	0
56	0	0	0	1	0
Всего	28	24	24	22	23



ОИЗ
УКЦ

10,000

7,500

5,000

2