ЭПИДЕМИОЛОГИЯ · ЦИФРОВОЕ ЗДРАВООХРАНЕНИЕ

Анализ больших данных в эпидемиологии

Прогнозирование вспышек заболеваний: данные, методы, кейсы, ограничения, внедрение

Презентация для трёх докладчиков · 2026

СТРУКТУРА

Три части — три докладчика

I

Данные и надзор

Откуда берутся данные, как устроен пайплайн, чем эпидемиологические данные отличаются от обычных.

Докладчик 1 · слайды 3–14

II

Методы и прогноз

Статистика, машинное обучение, механистические модели, ансамбли, реальные кейсы и провалы.

Докладчик 2 · слайды 15–31

III

Внедрение и риски

Ограничения, валидация, архитектура, дорожная карта, экономика и этика.

Докладчик 3 · слайды 32–45

Логика: данные → методы → внедрение. Каждая часть заканчивается тремя выводами.

ЧАСТЬ I

Данные и надзор

Какие данные есть у эпидемиолога, почему они грязные и как из них собирается система раннего предупреждения.

Докладчик 1

ПРОБЛЕМА

Классический надзор опаздывает — а вспышка не ждёт

- Пациент приходит не в первый день болезни: при туберкулёзе — через недели, при ОРВИ — на 2–3-й день.
- Врач должен заподозрить заболевание, назначить обследование и дождаться результата.
- Лаборатория подтверждает случай, форма уходит в СЭС, данные агрегируются в отчёт.
- Каждый этап добавляет дни: суммарно от события до цифры в отчёте проходит 1–3 недели.
- При гриппе и COVID-19 число случаев удваивается каждые 2–4 дня — прогноз устаревает раньше, чем выходит отчёт.
- Вмешательство, начатое по факту, стоит дороже и даёт меньший эффект, чем начатое по сигналу.

1–3 нед

типичное запаздывание отчётности

7–14 дней

рабочий горизонт прогноза

24/7

данные, которые уже есть в системах

ПРОБЛЕМА

Что даёт выигрыш во времени

- Раннее вмешательство действует на меньшем числе случаев — эффект выше при меньших затратах.
- Точный прогноз позволяет усилить конкретный район, а не закрывать весь регион.
- Изоляция и ограничения без данных превращаются в тотальные меры «на всякий случай».
- Логистика (вакцины, тесты, койки, бригады) планируется заранее, а не в режиме пожара.
- Общественность получает объяснимое основание для мер — это снижает сопротивление.
- Данные нужны не для отчёта, а для решения: где, когда и что делать.

Точечность

меры в конкретном очаге вместо всеобщих

Логистика

ресурсы едут туда, где будет рост

Доверие

решения объяснимы и проверяемы

ОПРЕДЕЛЕНИЕ

Что такое большие данные: пять V

Volume — объём

Миллионы записей ЭМК, миллиарды поисковых запросов, терабайты лабораторных результатов и геномных последовательностей.

Velocity — скорость

Поток в реальном времени: минуты и часы вместо месяцев. Именно скорость даёт выигрыш в эпидемиологии.

Variety — разнообразие

Текст, геоданные, изображения, геном, сигналы сенсоров и носимых устройств, данные аптек и сточных вод.

Veracity — достоверность

Шум, пропуски, дубли, ошибки кодирования, разная чувствительность тестов. Данные не собраны под задачу.

Value — ценность

Не объём, а пригодность для конкретного решения: чем раньше сигнал, тем он ценнее.

ОСОБЕННОСТЬ

Чем эпидемиологические данные отличаются от «обычных»

Источники неудобства

- Наблюдательные данные, а не эксперимент: мы не можем рандомизировать вспышку.
- Сильное смещение тестирования: учитываются те, кого тестировали, а не все больные.
- Задержка отчётности: последние дни всегда неполные, и это надо корректировать.
- Малые числа по районам: 3 случая — это ещё не вспышка, но и не ноль.
- Сезонность, дрейф и изменение политики тестирования меняют сам смысл показателя.

Что с этим делать

- Корректировать на задержку: nowcasting вместо «сырых» последних точек.
- Иерархические байесовские модели для малых территорий — заимствование силы.
- Добавлять внешние ковариаты: климат, мобильность, плотность населения.
- Постоянно контролировать дрейф и переобучать модель по расписанию.
- Фиксировать допущения письменно: прогноз без них невозможно проверить.

ИСТОЧНИКИ

Медицинские системы

Электронные медкарты

Жалобы, анамнез, диагнозы, назначения, даты обращения. Самый детальный, но самый медленный источник.

Лабораторные ИС

ПЦР, посевы, серология, антибиотикорезистентность. Дают подтверждение и характеристику возбудителя.

Регистры

Туберкулёз, ВИЧ, вакцинация, хронические болезни. Позволяют отслеживать когорты и исходы лечения.

Госпитальная статистика

Поступления, койко-дни, ОРИТ, летальность. Хороший индикатор тяжести и нагрузки на систему.

Аптечные продажи

Рост продаж жаропонижающих и противовирусных — косвенный сигнал роста заболеваемости.

Телемедицина и скорая

Обращения и вызовы: самый ранний медицинский сигнал, опережает диагноз на дни.

ИСТОЧНИКИ

Цифровые следы

Поисковые запросы

«температура», «кашель», «как сдать мокроту». Работают быстро, но связь с болезнью нестабильна.

Социальные сети

Упоминания симптомов, тревога, стигма, отказы от вакцинации, слухи. Требуют NLP и осторожности.

Мобильность

Перемещения между регионами, скопления, занос инфекции. Основа агентных моделей.

Геоданные

Плотность, транспортные потоки, доступность медицинской помощи.

Новости и отчёты

Вспышки в других странах, закрытие школ, сообщения о неизвестной пневмонии.

Онлайн-опросы

Самоотчёты о симптомах: добровольные когорты, быстрый охват, смещение выборки.

ИСТОЧНИКИ

Среда, биология и смертность

Сточные воды

РНК возбудителя и антибиотикорезистентность. Рост в стоках опережает госпитализации на 5–7 дней.

Геномика

Секвенирование линий возбудителя: вспышка, занос, устойчивость, связь случаев.

Климат и переносчики

Температура, осадки, ареалы комаров — основа прогноза денге, малярии, лихорадки Западного Нила.

Избыточная смертность

Разница с ожидаемым уровнем: показывает истинный масштаб, когда тестов мало.

Ветеринарный надзор

Зоонозные вспышки у животных — ранний признак риска для людей.

Экология и вода

Качество воды, санитария, доступ к ней — ключ к прогнозу холеры и кишечных инфекций.

СРАВНЕНИЕ

Классический и цифровой надзор

Критерий	Классический надзор	Цифровой надзор
Латентность	1–3 недели	часы — дни
Полнота	высокая: подтверждённые случаи	переменная, требует валидации
Детализация	агрегаты по территории	уровень пациента и адреса
Стоимость	высокая: ручной сбор и отчётность	ниже на единицу данных
Устойчивость	мало зависит от алгоритмов	чувствителен к смещению и дрейфу
Роль в системе	эталон истины	ранний сигнал и прогноз
Итог	надёжно, но поздно	быстро, но требует проверки

ПАЙПЛАЙН

От сырых данных до решения — шесть шагов

1. Сбор

API, HL7/FHIR, выгрузки из ИС, регистров и лабораторий. Фиксируем дату события и дату поступления записи.

2. Очистка

Дубли, пропуски, выбросы, единицы измерения, противоречивые даты. Здесь теряется больше всего времени.

3. Интеграция

Единый ID пациента, геокодирование адреса, согласование справочников и классификаторов.

4. Признаки

Лаги, скользящие средние, сезонность, тренд, кодирование дней недели и праздников.

5. Модель

Статистика, машинное обучение, механистические модели, ансамбль и калибровка.

6. Решение

Валидация, порог срабатывания, оповещение, действие и обратная связь по результату.

КАЧЕСТВО ДАННЫХ

Шесть типовых дефектов и как их ловить

Дефект	Пример	Что делать
Пропуски	нет результата теста у части случаев	множественное восстановление, пометка «неизвестно»
Дубли	один случай из двух лабораторий	дедупликация по ID, дате и ФИО
Ошибки кодирования	неверный код МКБ в ЭМК	сверка со справочником, ручная выборка
Сдвиг дат	дата регистрации вместо даты начала	разделять дату события и дату записи
Разная гранулярность	район против города	агрегация до общего уровня
Нет единого ID	пациент в двух системах	мастер-индекс пациентов

ИТОГ ЧАСТИ I

Три вывода по данным

- Данных много, но они грязные, запаздывают и смещены — это свойство предметной области, а не недоработка.
- Скорость важнее полноты: ранний неточный сигнал полезнее позднего точного отчёта.
- Качество пайплайна определяет потолок качества любой модели — сначала данные, потом алгоритмы.

ЧАСТЬ II

Методы и прогноз

Как из данных получается прогноз: постановка задачи, метрики, модели, реальные кейсы и типичные провалы.

Докладчик 2

ПОСТАНОВКА ЗАДАЧИ

Четыре типа задач прогнозирования

Классификация

Будет ли вспышка в регионе через 2 недели: да или нет. Метрики — precision, recall, AUC. Цена ошибки зависит от асимметрии.

Регрессия

Сколько случаев будет через 7, 14 или 28 дней. Метрики — MAE, MAPE, покрытие интервала.

Nowcasting

Сколько случаев уже есть, но ещё не учтено из-за задержки отчётности. Критично для текущих решений.

Сигнальный надзор

Есть ли аномалия в потоке данных. Метрики — время до обнаружения и доля ложных тревог.

МЕТРИКИ

Чем измерять прогноз и где метрика врёт

Метрика	Что показывает	Где врёт
AUC	различение «вспышка / нет»	при сильном дисбалансе классов
Precision / recall	точность и полнота сигналов	не учитывает время обнаружения
MAE / MAPE	ошибка числа случаев	MAPE взрывается на малых числах
Brier score	калибровку вероятностей	не видит форму кривой
Coverage	честность интервалов	легко завысить шириной
WIS	качество всего прогноза	сложно объяснить заказчику

МЕТОДЫ

Статистика и nowcasting

- SARIMA и экспоненциальное сглаживание — базовая линия, с которой сравнивают всё остальное.
- Байесовские иерархические модели дают устойчивые оценки для малых районов за счёт обмена информацией.
- Nowcasting корректирует неполные последние дни: оценивает задержку по историческим данным.
- R_t оценивают по дате заражения, а не по дате отчёта — иначе динамика искажается.
- Плюсы: интерпретируемость, интервалы, работа на коротких рядах.
- Минусы: плохо ловят смену режима и вмешательства без явного моделирования.

Машинное обучение: что реально работает

- Градиентный бустинг и случайный лес — лучший выбор для табличных данных с ковариатами.
- Лассо и отбор признаков помогают, когда переменных много, а случаев мало.
- Обнаружение аномалий (Isolation Forest, автоэнкодеры) — для сигнального надзора.
- Ключ к результату — признаки: лаги, скользящие окна, сезонность, внешние ковариаты.
- Ограничения: мало данных, дрейф, риск переобучения на шум и утечки.
- ML не объясняет механизм — поэтому его дополняют механистическими моделями.

МЕТОДЫ

Глубокое обучение и NLP

- LSTM и трансформеры работают на длинных рядах и множестве территорий одновременно.
- NLP обрабатывает новости, соцсети и жалобы: симптомы, слухи, стигма, отказы от вакцинации.
- Мультимодальные модели объединяют ряды, текст и геоданные.
- Оправданы при больших объёмах и сложной структуре; иначе проигрывают простым моделям.
- Требуют инфраструктуры, специалистов и осторожной валидации.
- Интерпретация затруднена — нужны SHAP, attention-анализ и участие эпидемиолога.

Механистические модели: SIR, SEIR, агентные

- SIR/SEIR описывают переходы: восприимчивые → инфицированные → выздоровевшие.
- Возрастная структура и социальные контакты объясняют разную скорость распространения.
- Агентные модели воспроизводят город: школы, транспорт, работу, домохозяйства.
- Позволяют оценивать сценарии вмешательств: маски, дистанция, вакцинация, закрытие школ.
- Сильны смыслом и структурой, слабы в точных числах без калибровки на реальные данные.
- Гибрид с ML даёт и структуру, и поправку на фактическую динамику.

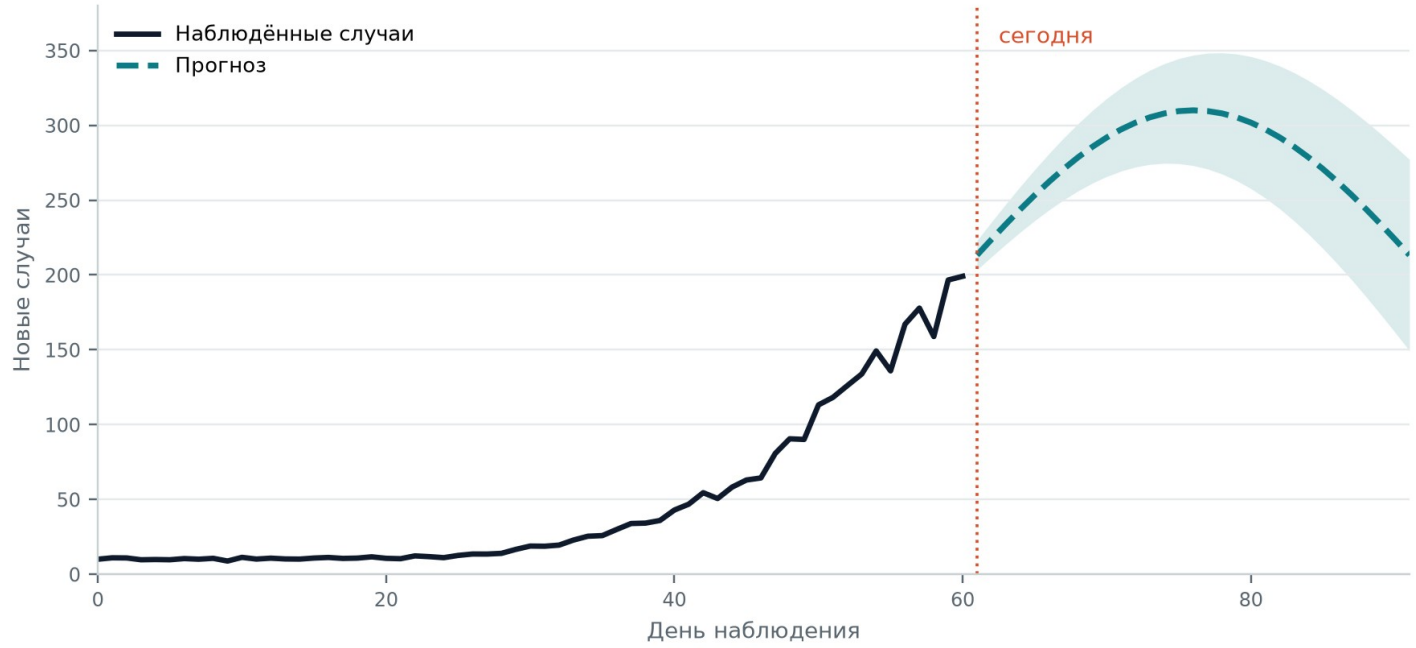
МЕТОДЫ

Ансамбли и гибриды

- Ансамбль объединяет прогнозы разных моделей — обычно он точнее и стабильнее каждой по отдельности.
- FluSight и COVID-19 Forecast Hub показали: ансамбль редко бывает хуже средней модели.
- Веса моделей подбирают по исторической точности, а не по красоте алгоритма.
- Гибрид: механистическая структура + ML-поправка на реальные данные и поведение.
- Обязательна калибровка интервалов — иначе ансамбль даёт ложную уверенность.
- Ансамбль не спасает от плохих данных: мусор на входе остаётся мусором.

ИЛЛЮСТРАЦИЯ

Прогноз — это число, интервал и горизонт



- Сплошная линия — наблюдённые случаи, пунктир — прогноз.
- Закрашенная зона — 80 % интервал неопределённости.
- Чем длиннее горизонт, тем шире интервал — это честно, а не дефект.
- Вертикаль «сегодня» отделяет факт от прогноза.
- **Прогноз без интервала — не прогноз, а догадка.**

Данные синтетические — показана форма прогноза, а не реальная вспышка.

Чек-лист перед запуском

Есть базовая линия?

Сравните с наивным прогнозом «как вчера» и с SARIMA. Если разницы нет — модель не нужна.

Хватает данных?

Нужны минимум 2–3 сезона и понятная структура. Иначе — простая модель и широкий интервал.

Понятна цена ошибки?

Ложная тревога и пропущенная вспышка стоят по-разному — порог выбирается под это.

Можно объяснить?

Эпидемиолог должен понимать, какие признаки двигают прогноз.

Есть план валидации?

Разделение по времени, backtesting, проспективная проверка.

Кто принимает решение?

Модель предлагает, человек отвечает — это фиксируется регламентом.

КЕЙС

Грипп: FluSight и ансамблевый прогноз

- CDC собирает прогнозы десятков команд и публикует ансамбль каждую неделю.
- Цель — госпитализации на 1–4 недели вперёд по регионам.
- Ансамбль стабильно точнее средней отдельной модели и реже даёт выбросы.
- Качество оценивают по WIS и покрытию интервалов, а не по одному числу.
- Сезонность и циркулирующие штаммы — ключевые признаки.
- Результат используется для планирования коек и вакцин, а не для публичных сенсаций.

КЕЙС

Лихорадка денге: климат плюс переносчик

- Заболеваемость зависит от температуры, осадков и численности Aedes.
- Модели объединяют климатические прогнозы, энтомологические данные и мобильность.
- Горизонт — 1–3 месяца, что достаточно для планирования обработок и закупок.
- Раннее предупреждение позволяет начать борьбу с переносчиком до роста случаев.
- Ограничение: качество энтомологических данных и нелинейная связь с климатом.
- Кейс показывает ценность средовых данных, которые не связаны с медициной напрямую.

КЕЙС

Холера в Йемене, 2017–2019

- Крупнейшая вспышка последних десятилетий: сотни тысяч случаев, разрушенная инфраструктура.
- Прогноз строился на метаданных, спутниковых данных и качестве воды.
- Горизонт — 4 недели: достаточно для распределения вакцин и организации водоснабжения.
- Модель учитывала плотность населения, сезонность и доступ к чистой воде.
- Ограничение: данные о случаях неполные, часть больных не обращается.
- Итог: прогноз встроен в гуманитарное планирование, а не в академическую статью.

КЕЙС

COVID-19: nowcasting, R_t и сточные воды

- Nowcasting корректировал неполные последние дни по задержке отчётности.
- R_t оценивали по дате заражения, а не по дате регистрации.
- Ансамбли Forecast Hub давали недельные прогнозы госпитализаций по странам и регионам.
- Сточные воды давали сигнал за 5–7 дней до роста госпитализаций и не зависели от тестирования.
- Главный урок: смена политики тестирования ломала модели, обученные на прошлых данных.
- Успех был там, где прогноз обновлялся часто и проверялся проспективно.

ПРОВАЛ

Google Flu Trends: как не надо делать

Что произошло

- 2008: прогноз гриппа по поисковым запросам опережал CDC на 1–2 недели.
- 2013: на пике сезона ошибка почти двукратная.
- 2015: проект закрыт.
- Причина — дрейф связи «запрос — болезнь» и отсутствие постоянной сверки с эталоном.

Уроки

- Корреляции в цифровых следах нестабильны: нужен постоянный контроль.
- Ретроспективные метрики с утечкой данных не гарантируют пользы.
- Модель без эпидемиологического смысла ломается при смене поведения.
- Публичный прогноз без интервала и разбора ошибок убивает доверие.

Услышать вспышку раньше, чем её посчитают

- Event-based: новости, отчёты, соцсети на десятках языков — ProMED, HealthMap, GPHIN, BlueDot.
- BlueDot сообщил о кластере пневмонии в Ухане 31 декабря 2019 года — раньше официальных сообщений.
- Indicator-based: официальная отчётность — медленнее, но точнее.
- Рабочая схема: сигнал → верификация эпидемиологом → решение.
- Метрики: время до обнаружения, доля ложных тревог, полнота охвата источников.
- Сигнал не равен диагнозу: ценность в скорости, точность добавляет проверка.

ИТОГ ЧАСТИ II

Три вывода по методам

1 **Модель — не цель**

Цель — решение; модель лишь оценивает вероятность и масштаб.

2 **Базовая линия обязательна**

Без сравнения с наивным прогнозом сложность не оправдана.

3 **Ансамбль и калибровка**

Несколько моделей и честные интервалы надёжнее одной «лучшей».

ЧАСТЬ III

Внедрение и риски

Где модели ломаются, как их проверять, какая нужна инфраструктура и что делать в Республике Казахстан.

Докладчик 3

ОГРАНИЧЕНИЯ

Где ломаются данные

Смещение тестирования

Тестируют не всех: доступность тестов и решения врачей искажают картину сильнее, чем сама эпидемия.

Смена политики

Изменение правил тестирования выглядит как вспышка или как её окончание.

Малые числа

Три случая в районе — это шум; модель должна это признавать, а не строить тренд.

Пропуски и дубли

Один пациент в двух системах и отсутствие результата у части случаев.

Задержка отчётности

Последние дни всегда неполные — без коррекции прогноз систематически занижен.

Разные справочники

Коды МКБ, названия районов и единицы измерения не совпадают между системами.

ОГРАНИЧЕНИЯ

Где ломаются модели

- Переобучение: модель запоминает шум сезона и плохо работает на новом.
- Дрейф концепции: связь признаков с исходом меняется со временем.
- Утечка данных: признаки, которые становятся известны только после события.
- «Чёрный ящик»: прогноз невозможно объяснить, а значит — защитить перед руководством.
- Обратная связь: публичный прогноз меняет поведение людей и тем самым исход.
- Нестабильность: малые изменения данных дают большие изменения прогноза.

Приватность, согласие, стигма

- Медицинские данные — особо чувствительная категория: нужны правовые основания обработки.
- Геоданные легко деанонимизируют человека, даже без имени и фамилии.
- Цели использования фиксируются: нельзя использовать данные для дискриминации.
- Стигматизация групп риска (мигранты, заключённые, ЛЖВ) усиливает сокрытие случаев.
- Цифровое неравенство: в слабых регионах меньше данных — и хуже прогноз.
- Ответственность за решение остаётся за человеком, а не за моделью.

ВАЛИДАЦИЯ

Как проверить прогноз, а не поверить ему

- Разделение по времени, а не случайное: будущее нельзя подсматривать.
- Backtesting на 3–5 прошлых сезонах — обязательный минимум.
- Проспективная валидация: модель работает «вслепую» до наступления сезона.
- Внешняя проверка на другом регионе или периоде.
- Метрики: различение (AUC), калибровка (Brier), точность (MAE), честность (coverage).
- Решение принимает человек: модель предлагает, эпидемиолог подтверждает.

МОНИТОРИНГ

Что делать после внедрения

- Отслеживать дрейф данных и деградацию метрик на новых периодах.
- Переобучать по расписанию, а не только при аварии.
- Версионировать модели, данные и признаки — иначе результат невоспроизводим.
- Журнал решений: какой прогноз, кто и на основании чего принял меры.
- Алерты на качество данных: пропал поток, изменился справочник, выросли пропуски.
- Разбор ошибок после каждой вспышки — обязательная часть цикла.

АРХИТЕКТУРА

Как это выглядит в системе здравоохранения

Источники	ЕИСЗ и ЭМК · лабораторные ИС · регистры туберкулёза и ВИЧ · отчётность СЭС · аптеки · сточные воды
Хранилище	data lake для сырых данных · хранилище очищенных · каталог данных и словарь показателей
Обработка	ETL/ELT · потоковая обработка · контроль качества · мастер-индекс пациентов и геокодирование
Аналитика	feature store · реестр моделей · обучение и валидация · мониторинг дрейфа и деградации
Потребление	дашборды для СЭС и ПМСП · алерты · API для регионов · отчёты для руководства
Управление	разграничение доступа · аудит · обезличивание · соответствие Кодексу о здоровье

Что уже есть и чего не хватает

- Единая информационная система здравоохранения и электронные медкарты накапливают данные ПМСП.
- Работают регистры по туберкулёзу и ВИЧ, отчётность санитарно-эпидемиологической службы.
- Есть система лабораторной диагностики и данные о вакцинации.
- Барьер 1: разрозненность систем и разные справочники.
- Барьер 2: качество кодирования и отсутствие единого идентификатора пациента.
- Барьер 3: нехватка аналитиков и эпидемиологов с навыками работы с данными.

ДОРОЖНАЯ КАРТА

Два года до работающей системы

0–3 месяца

Аудит данных, словарь показателей, пилот в одном регионе, эталон для сверки, команда проекта.

3–6 месяцев

Пайплайн, nowcasting обращаемости, первый дашборд, обучение эпидемиологов.

6–12 месяцев

Прогноз на 7–14 дней с интервалами, ансамбль, алерты, проспективная валидация.

12–24 месяца

Масштабирование, сточные воды и геномика, интеграция с ЛПУ, регламент решений.

РОЛИ

Кто что делает

Роль	Задача	Результат
Эпидемиолог	формулирует вопрос, проверяет сигнал	корректная постановка и интерпретация
Аналитик данных	пайплайн, модели, метрики	работающий прогноз и отчёты
ИТ-служба	интеграции, доступы, инфраструктура	устойчивый поток данных
ПМСП и фтизиатр	полнота и качество первичных данных	достоверная основа прогноза
Руководство	регламент и ответственность за решения	решения на основе прогноза

РИСКИ

Риски внедрения и меры

Риск	Проявление	Мера
Плохие данные	прогноз не сходится с фактом	контроль качества и эталон
Недоверие врачей	данные вносят формально	обучение и обратная связь
Утечка ПДн	деанонимизация пациентов	обезличивание и аудит доступа
Ложные тревоги	алерты игнорируют	настройка порогов и разбор
Нет владельца	проект умирает после пилота	регламент и ответственный
Дрейф модели	точность падает незаметно	мониторинг и переобучение

ЭКОНОМИКА

Стоимость и эффект

- Затраты: инфраструктура, интеграции, аналитики, обучение, сопровождение.
- Эффект: более ранняя реакция, меньше случаев, точечные меры вместо всеобщих.
- Экономия на масштабе реагирования обычно превышает стоимость аналитики.
- Считать нужно не стоимость модели, а стоимость решения, принятого без неё.
- Пилот на одном регионе дешевле и безопаснее, чем общенациональный запуск.
- Без регламента использования любой прогноз остаётся аналитическим отчётом.

ВЫВОДЫ

Пять тезисов на защиту

1	**Скорость, а не объём**	Big Data сокращает разрыв между событием и решением.
2	**Данные важнее алгоритма**	80 % результата — сбор, очистка и интеграция.
3	**Прогноз = число + интервал**	Без горизонта и оценки неопределённости он бесполезен.
4	**Валидация по времени**	Ретроспектива с утечкой данных — самообман.
5	**Организационная задача**	Регламент, обучение, ответственность и этика.

Данные есть. Вопрос только в том, успеем ли мы их использовать до вспышки.

Спасибо за внимание

Вопросы и обсуждение