Технология информационного анализа электрокардиосигналов

проф., д.м.н. Успенский Вячеслав Максимилианович,
проф. РАН, д.ф.-м.н. Воронцов Константин Вячеславович,
асп. МФТИ Бунакова Влада Руслановна,
асп. ФИЦ ИУ РАН Ишкина Шаура Хабировна

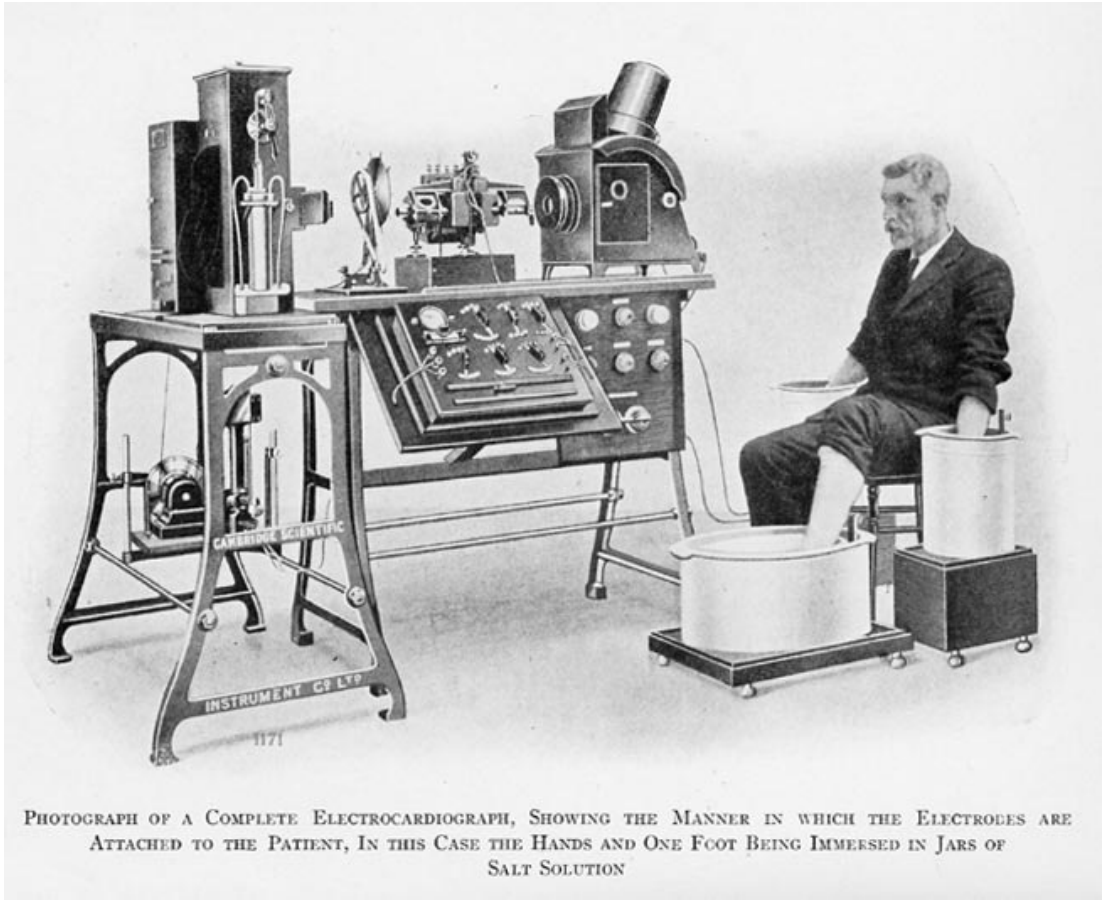
ВЦ РАН ФИЦ ИУ РАН • МФТИ

Гаджеты в медицине: перспективы использования, новые функции, техническое совершенствование, производство и финансирование

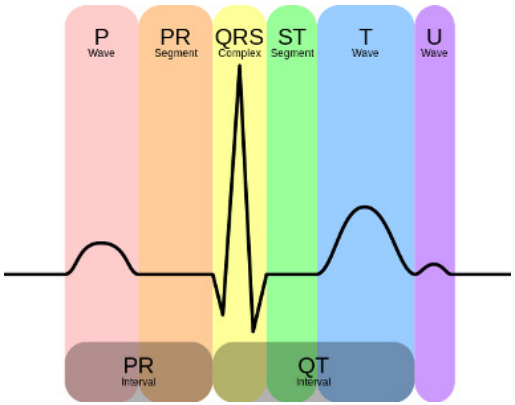
МГТУ имени Н.Э.Баумана • 13 марта 2017

Содержание

Электрокардиография



Photograph of a Complete Electrocardiograph, Showing the Manner in which the Electrodes are Attached to the Patient, In this Case the Hands and One Foot Being Immersed in Jars of Salt Solution



1872 — первые записи электрической активности сердца

1911 — коммерческий электрокардиограф (фото)

1924 — нобелевская премия по медицине, Виллем Эйнтховен

Теория информационной функции сердца [В.М.Успенский]

Возможна ли диагностика несердечных заболеваний по ЭКГ?

Предпосылки:

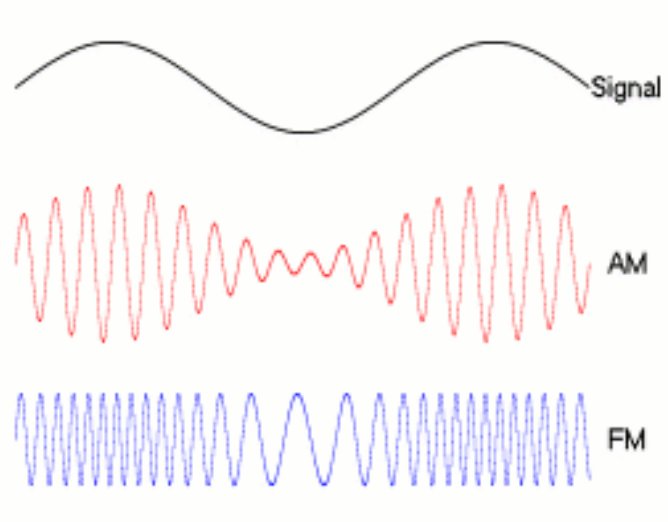
- Китайская традиционная медицина: *пульсовая диагностика*
- Р. М. Баевский: использование вариабельности сердечного ритма (*интервалов кардиоциклов*) в целях диагностики
- Цифровая электрокардиография высокого разрешения

Предположения:

- ЭКГ-сигнал несёт информацию о функционировании всех систем организма, не только сердца
- Информация о заболевании может проявляться на любой его стадии, поэтому возможна *ранняя* диагностика
- Каждое заболевание по-своему «модулирует» ЭКГ-сигнал

Аналогии в теорию передачи сигналов

Модуляция — процесс, при котором высокочастотная волна используется для переноса низкочастотного сигнала.



— низкочастотный сигнал

— амплитудная модуляция

— частотная модуляция

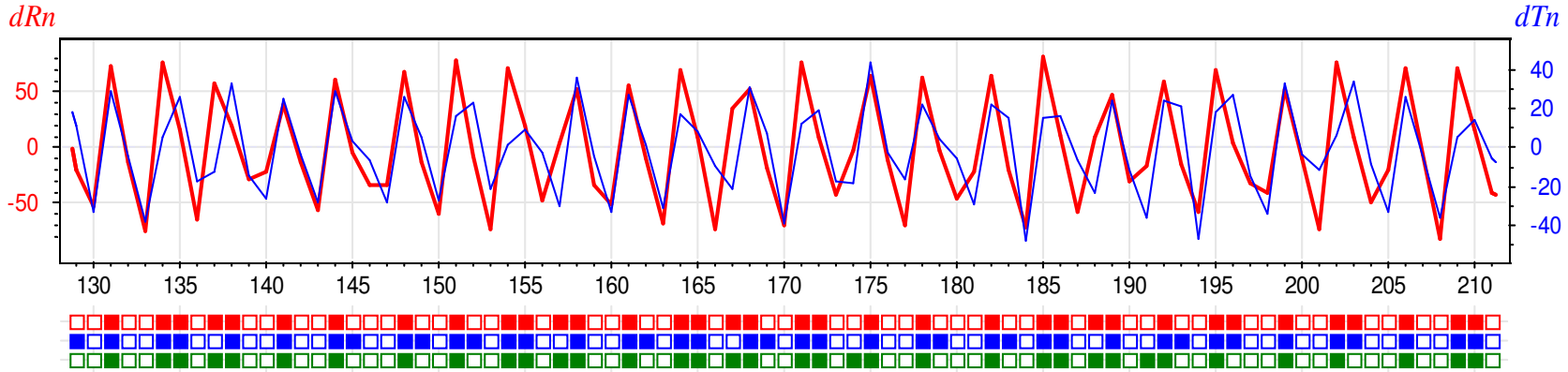
Демодуляция — процесс, обратный модуляции, преобразование модулированных колебаний высокой (несущей) частоты в исходный низкочастотный сигнал.

В случае ЭКГ несущая частота — биения сердца, ~ 1 Гц
А что будет аналогом модуляции и демодуляции?

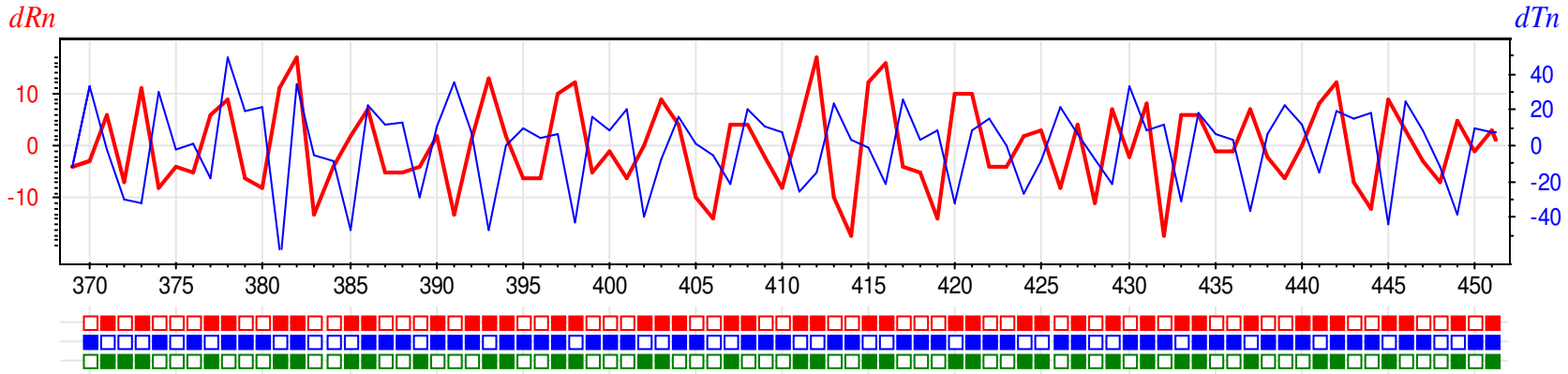
Есть ли различия в знаках приращений у больных и здоровых?

Приращения dR_t, dT_t, $d\alpha_t$ в последовательных кардиоциклах t

Здоровый:



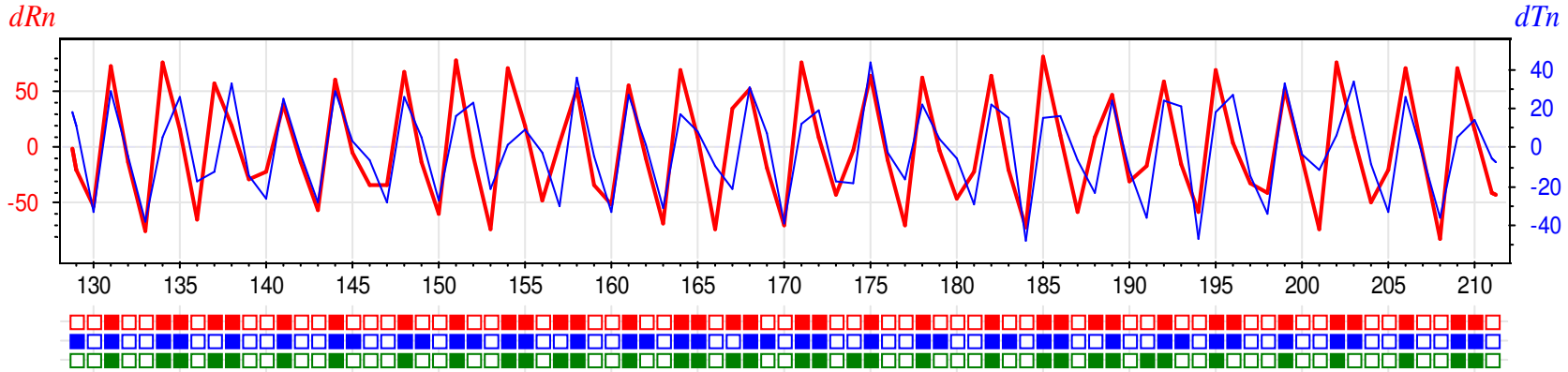
Больной (язвенная болезнь):



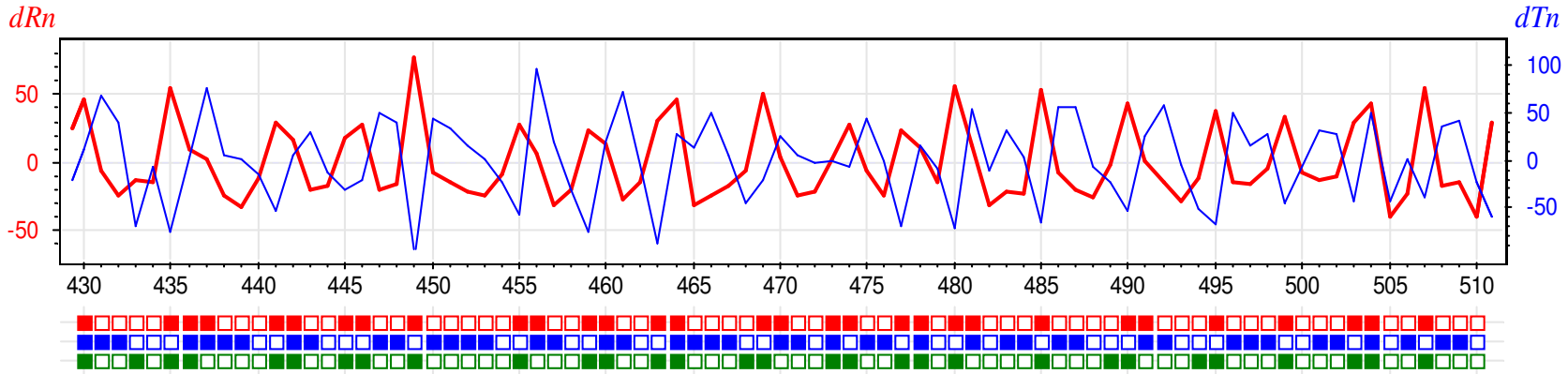
Есть ли различия в знаках приращений у больных и здоровых?

Приращения dR_t, dT_t, $d\alpha_t$ в последовательных кардиоциклах t

Здоровый:



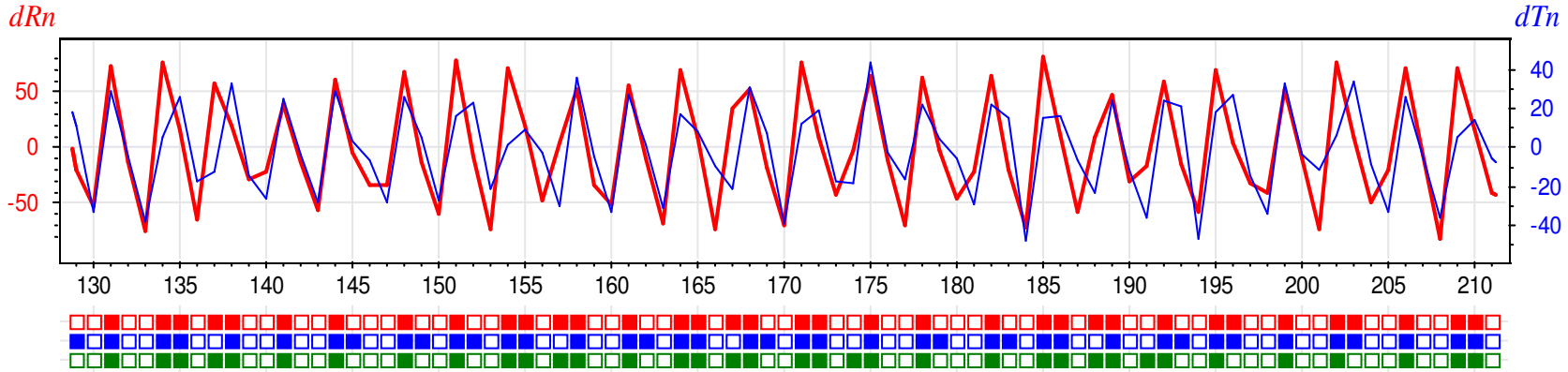
Больной (гипертония):



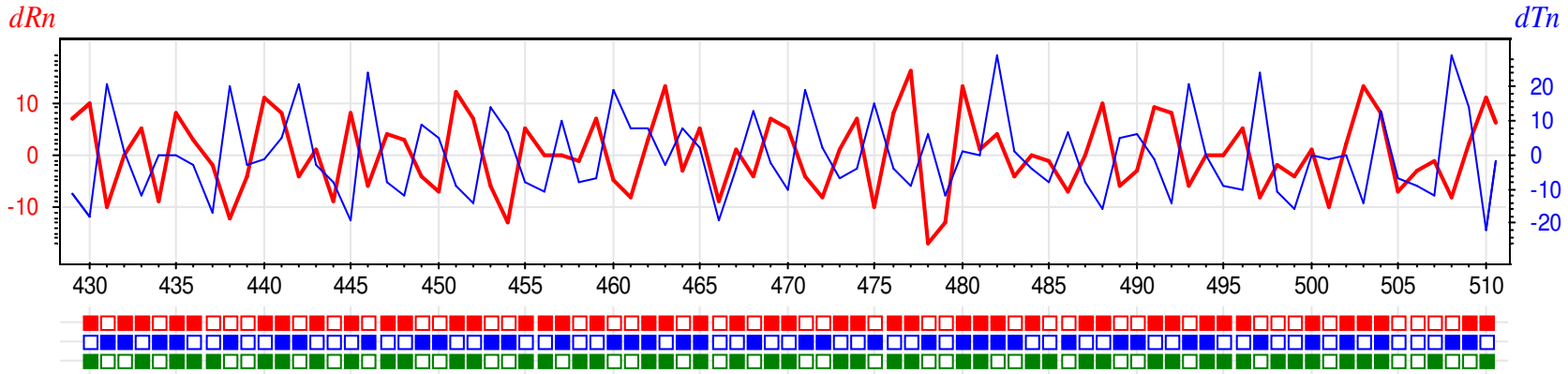
Есть ли различия в знаках приращений у больных и здоровых?

Приращения dR_t, dT_t, $d\alpha_t$ в последовательных кардиоциклах t

Здоровый:

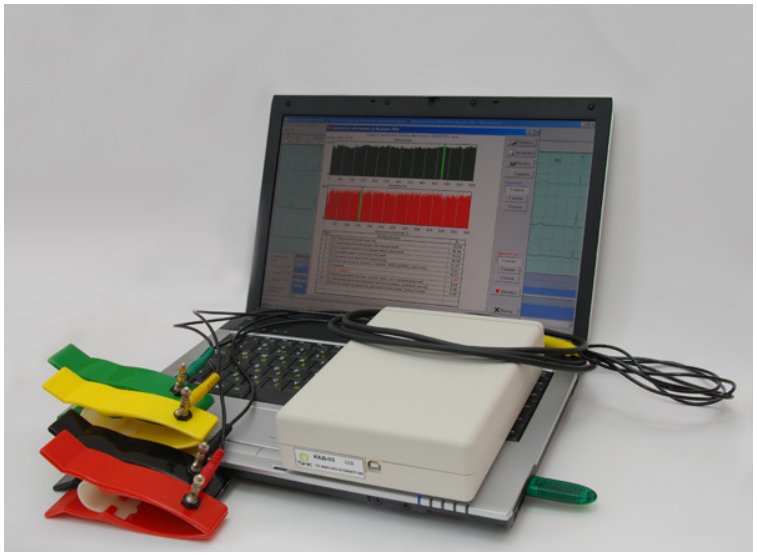


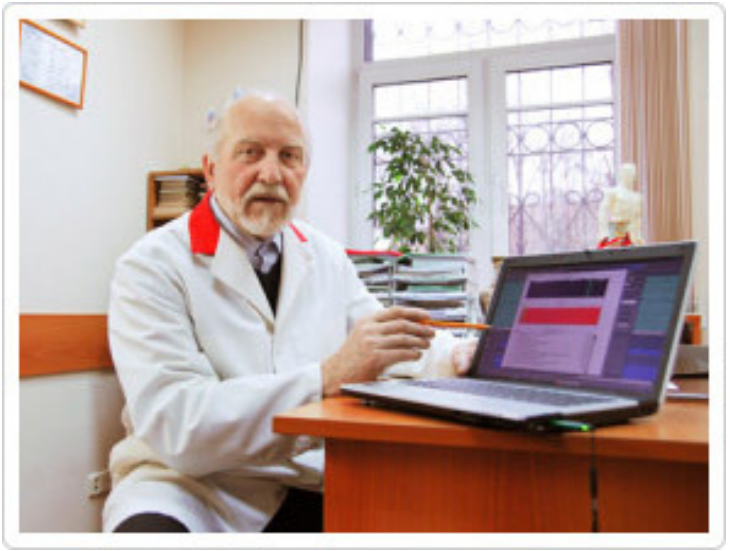
Больной (рак):



Диагностическая система «Скринфакс»

Цифровой электрокардиограф с улучшенной помехозащищённостью и расширенной полосой пропускания.





- более 15 лет исследований и накопления данных
- более 20 тысяч прецедентов (кардиограмма + диагнозы)
- более 40 заболеваний

Объём исходных данных (по заболеваниям)

«абсолютно здоровые»	АЗ	193
гипертоническая болезнь	ГБ	1894
ишемическая болезнь сердца	ИБС	1265
сахарный диабет (СД1 и СД2)	СД	871
язвенная болезнь	ЯБ	785
миома матки	ММ	781
узловой (диффузный) зоб щитовидный железы	УЩ	748
дискинезия желчевыводящих путей	ДЖВП	717
хронический гастрит (гастродуоденит) гипоацидный	ХГ2	700
вегетососудистая дистония	ВСД	694
мочекаменная болезнь	МКБ	654
рак общий (онкопатология различной локализации)	РО	530
холецистит хронический	ХХ	340
асептический некроз головки бедренной кости	НГБК	324
хронический гастрит (гастродуоденит) гиперацидный	ХГ1	324
желчнокаменная болезнь	ЖКБ	278
аднексит хронический	АХ	276
аденома простаты	ДГПЖ	260
анемия железодефицитная	ЖДА	260

Технология информационного анализа ЭКГ-сигналов

Этап I. Методы символьной динамики

1. *Демодуляция* — вычисление амплитуд, интервалов и углов по кардиограмме длиной 600 кардиоциклов
2. *Дискретизация* — перевод в *кодограмму* — 599-символьную строку в 6-буквенном алфавите
3. *Векторизация* — перевод в вектор 6^3=216 частот триграмм

Этап II. Методы машинного обучения

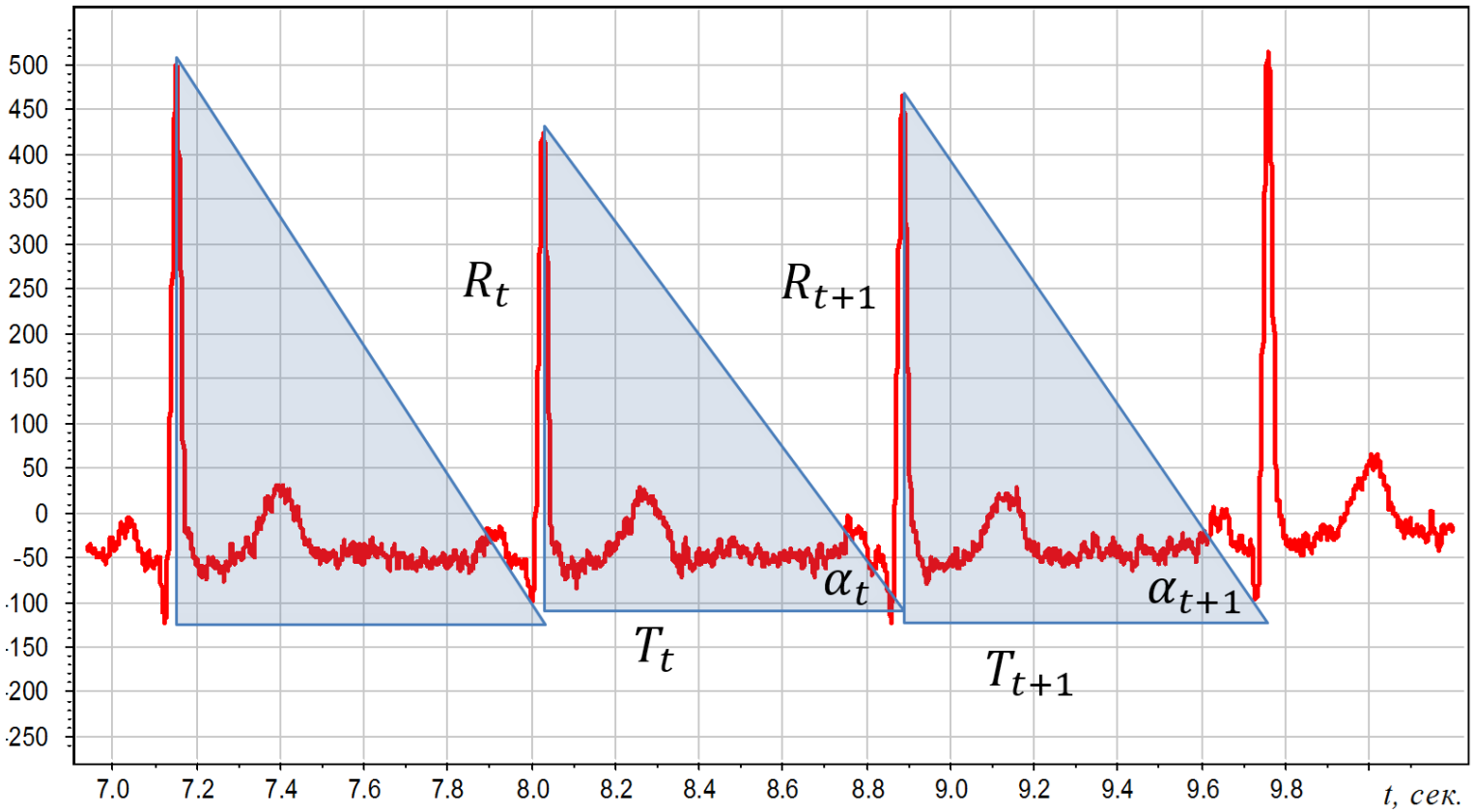
1. Формирование обучающих выборок здоровых и больных
2. Формировании модели классификации
3. Оптимизация модели классификации
4. Оценивание качества диагностики

Вариабельность интервалов и амплитуд кардиоциклов

приращение амплитуд: $dR_t = R_{t+1} - R_t$

приращение интервалов: $dT_t = T_{t+1} - T_t$

приращение углов: $d\alpha_t = \alpha_{t+1} - \alpha_t, \qquad \alpha_t = \operatorname{arctg} \frac{R_t}{T_t}$

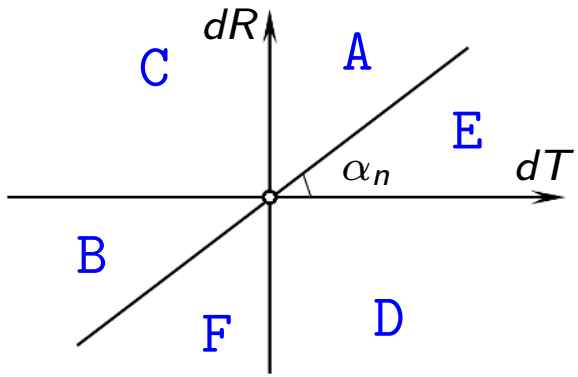


Дискретизация ЭКГ-сигнала

Вход: последовательность интервалов и амплитуд $(T_t, R_t)_{t=1}^{N+1}$

Правила кодирования:

$dR_t = R_{t+1} - R_t$	+	−	+	−	+	−
$dT_t = T_{t+1} - T_t$	+	−	−	+	+	−
$d\alpha_t = \alpha_{t+1} - \alpha_t$	+	+	+	−	−	−
s_t	A	B	C	D	E	F



Выход: *кодограмма* $x = (s_t)_{t=1}^{N}$ — последовательность символов алфавита $\{\mathtt{A}, \mathtt{B}, \mathtt{C}, \mathtt{D}, \mathtt{E}, \mathtt{F}\}$:

DBFEACFDAAFBABDDAADFAAFFEACFEACFBAEFFAABFFAAFFAAFFAAFFAAEBFAEBFEAAFCAAFFAAD
FCAFFAADFCADFCCDFDACFFACDFAEFFACFFEADFCAFBCADFFECFFAAFFAAFFAEFFCACFCAEFFCAD
DAADBFAAFFAEBFAABFACDFFAAFBAADFAADFDAAFCECFCEDFCEEFCAEFBECBBBAADBAACFFAAFFA
CFFCECFDAABDAEFFAAFFCEDBFAAFFAEFFAEFBACFBAEDFEAAFFCAFFDAAFFAEBDAADBBADFDAFF
EABFCCAFDEEBDECFFACFFAABFAADFBAAFFACFFFAEFFACFFACFFCECFBAAFFFAAFFFAAFFAADFB
AABFACDFDAEFFAADBAAEFFEAFBCECFDECCFBAAFFAADFDACDFAAFFAADFCAADFAEFBAAFFCADFE
AFFCECFCECFFAAFFABCFDAAAFFADBFCAEFFAABFACBFAAEBFAEBFCAFFBAAFFAAFFDACFDAABFB
CAFFAECFFACFFACDFCADFDAABFAAEDDABBFCACDBAAFFAAFFCADFAADFDACFFAEDFCACFCAEBCE

Векторизация ЭКГ-сигнала

По ЭКГ строится текстовая строка — *кодограмма*:

Частоты триграмм — число вхождений триграммы в кодограмму:

1. FFA - 42	17. EFF - 10	33. CEC - 6	49. EAC - 3
2. FAA - 33	18. DAA - 10	34. ADB - 5	50. DDA - 3
3. AFF - 32	19. ECF - 9	35. FFE - 5	51. CAC - 3
4. AAF - 30	20. FFC - 9	36. EBF - 5	52. EDF - 3
5. ADF - 18	21. FEA - 9	37. CFD - 5	53. EFB - 3
6. FCA - 18	22. DFC - 8	38. AFB - 4	54. DBA - 3
7. ACF - 17	23. ABF - 8	39. AAE - 4	55. FCC - 2
8. AAD - 15	24. AAB - 8	40. CFC - 4	56. AFC - 2
9. CFF - 14	25. FCE - 8	41. CAE - 4	57. EAA - 2
10. AEF - 13	26. AEB - 7	42. DAC - 4	58. CED - 2
11. FDA - 13	27. DFD - 7	43. DBF - 4	59. CAA - 2
12. FAE - 12	28. ACD - 6	44. BFC - 4	60. BCA - 2
13. FAC - 12	29. CDF - 6	45. CFB - 4	61. BBA - 2
14. FBA - 11	30. DFA - 6	46. AED - 3	62. DFF - 2
15. BFA - 11	31. CAF - 6	47. FFF - 3	63. BDA - 2
16. BAA - 11	32. CAD - 6	48. FBC - 3	64. DAE - 2

Линейная модель классификации с двумя классами

Для простоты рассмотрим только одно заболевание:
$y_i = 1$ — больной, $y_i = 0$ — здоровый

- чем выше частота триграммы x^j, тем она информативнее
- есть триграммы, более характерные для больных, и есть триграммы, более характерные для здоровых

Линейная модель классификации:

$$\langle x, w \rangle = \sum_{j=1}^{n} w_j x^j, \qquad a(x) = \begin{cases} 1, & \langle x, w \rangle \geqslant w_0 \\ 0, & \langle x, w \rangle < w_0 \end{cases}$$

где w_j — вес j-й триграммы:

- $w_j > 0$, если триграмма более характерна для больных
- $w_j < 0$, если триграмма более характерна для здоровых
- $w_j = 0$, если триграмма не информативна для этой болезни

Методы машинного обучения

- **Наивный байесовский классификатор**
  - ☺ простой интерпретируемый линейный классификатор
  - ☹ качество классификации невысокое
- **Наивный байесовский классификатор + отбор признаков**
  - 😐 качество классификации получше
  - 😐 находит один диагностический эталон каждой болезни
- **Метод главных компонент + логистическая регрессия**
  - ☺ качество классификации высокое
  - ☹ не определяет диагностические эталоны болезней
- **SVM, нейронные сети, случайный лес**
  - ☺ качество классификации высокое
  - ☹ неоправданно сложное, неинтерпретируемое решение
- **Тематические модели классификации**
  - ☺ автоматически находит все диагностические эталоны
  - 😐 качество классификации среднее

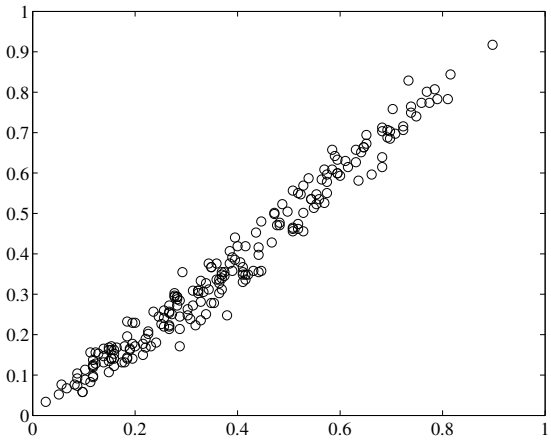
Существуют сочетания триграмм, специфичные для болезней

Точки на графиках соответствуют триграммам, $j = 1, \ldots, 216$

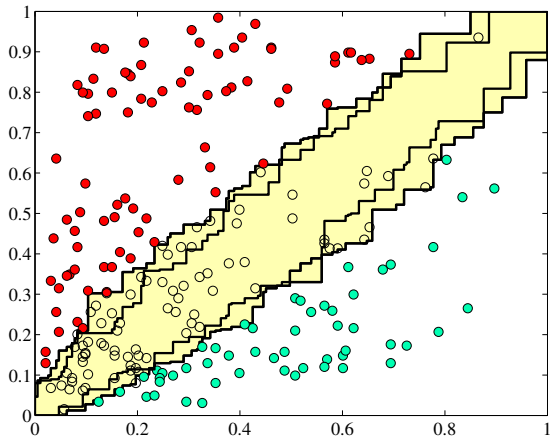
— ось X: доля здоровых x_i с частотой триграммы $x_i^j \geqslant 2$ из 600

— ось Y: доля больных x_i с частотой триграммы $x_i^j \geqslant 2$ из 600

НГБК (асептический некроз головки бедренной кости)



случайно перемешанные y_i



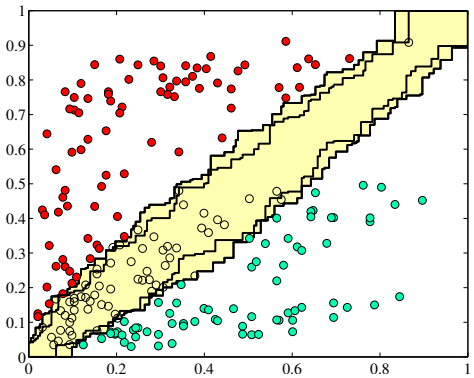
наблюдаемые y_i

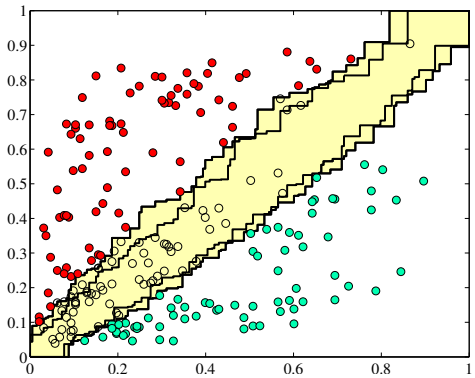
Слева: как распределятся точки, если объектам x_i назначить случайные (случайно перемешанные) метки классов y_i.

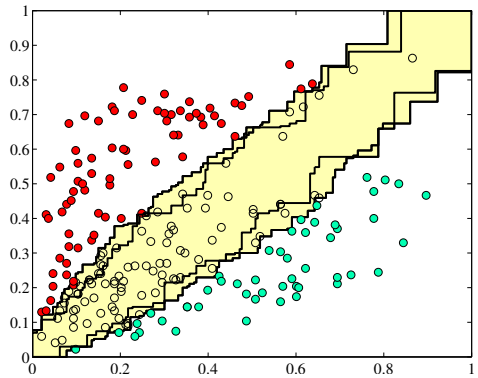
Жёлтая область: если случайно перемешать 20 раз, 1000 раз.

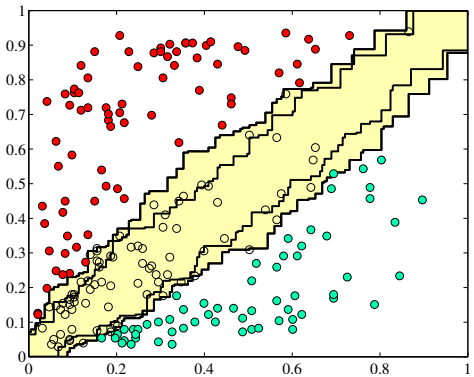
Существуют сочетания триграмм, специфичные для болезней

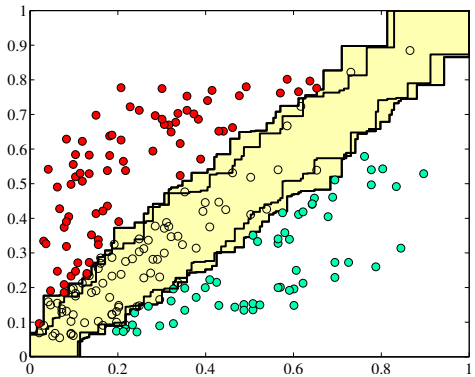
Для каждой болезни есть свои неслучайно частые триграммы

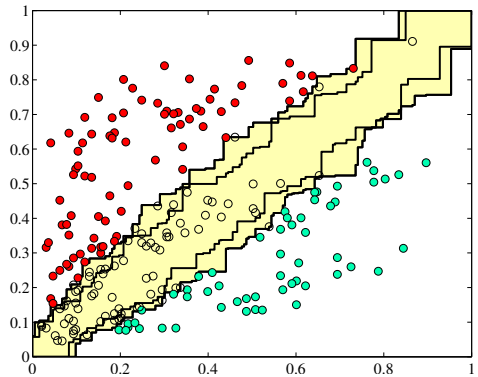

ишемия сердца


гипертония


рак

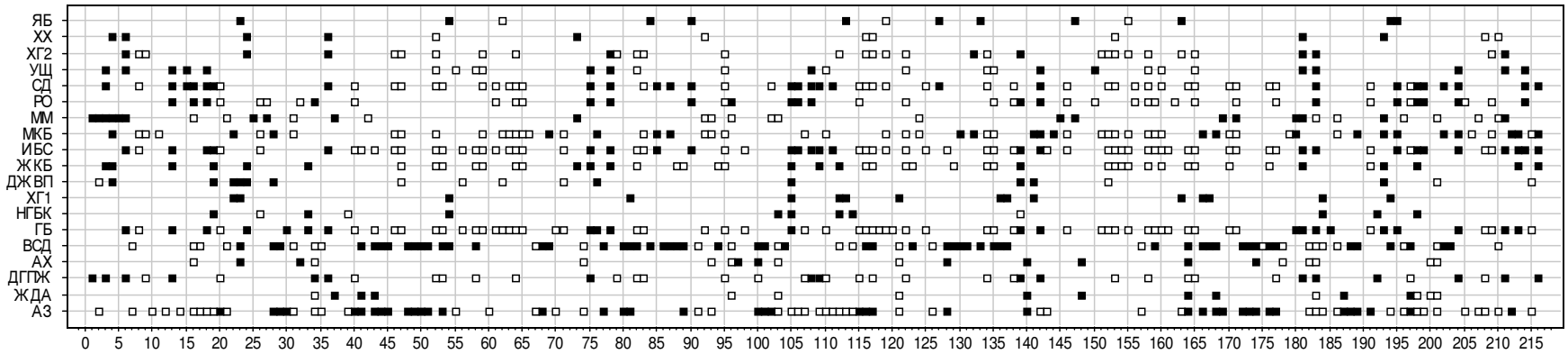

желчнокаменная болезнь


миома матки


язвенная болезнь

Болезни отличаются наборами информативных триграмм

ось X: — номера триграмм $j = 1, \ldots, n$, $n = 216$
ось Y: болезни (АЗ — абсолютно здоровые)



□ — неслучайно низкая частота триграммы
■ — неслучайно высокая частота триграммы

Вывод 1. Для каждой болезни есть триграммы с неслучайно высокой и неслучайно низкой частотой

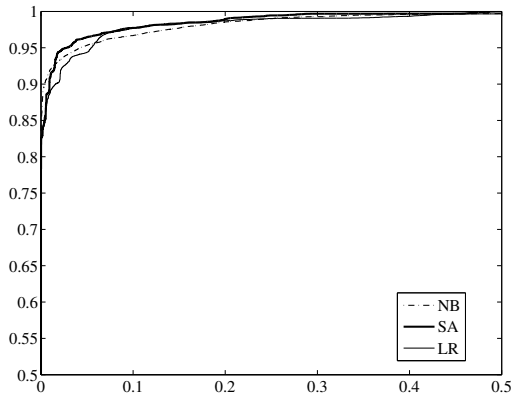
Вывод 2. *Диагностический эталон* болезни — специфичное подмножество триграмм с неслучайно высокой частотой

Результаты кросс-валидации

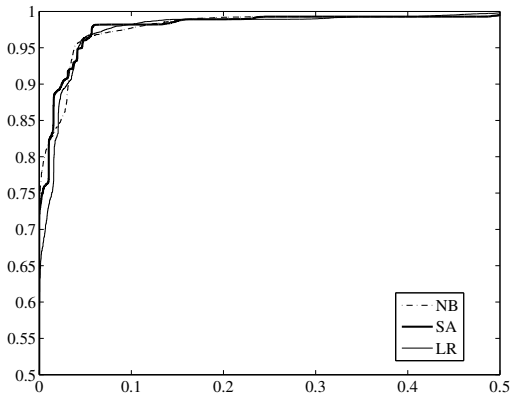
Обучающая выборка: оптимизация параметров модели
Тестовая выборка: Чувствительность, Специфичность, AUC
40×10-fold cross-validation — для доверительного оценивания

болезнь	выборка	AUC, %	С% при Ч=95%
некроз головки бедренной кости	327	99.19 ± 0.10	96.6 ± 1.76
желчнокаменная болезнь	277	98.98 ± 0.23	94.4 ± 1.54
ишемическая болезнь сердца	1262	97.98 ± 0.14	91.1 ± 1.86
гастрит	321	97.76 ± 0.11	88.3 ± 2.64
гипертоническая болезнь	1891	96.76 ± 0.09	84.7 ± 1.99
сахарный диабет	868	96.75 ± 0.19	85.3 ± 2.18
аденома простаты	257	96.49 ± 0.13	80.1 ± 3.19
рак	525	96.49 ± 0.28	82.2 ± 2.38
узловой зоб щитовидной железы	750	95.57 ± 0.16	73.5 ± 3.41
холецистит хронический	336	95.35 ± 0.12	74.8 ± 2.46
дискинезия ЖВП	714	94.99 ± 0.16	70.3 ± 4.67
мочекаменная болезнь	649	94.99 ± 0.11	69.3 ± 2.14
язвенная болезнь	779	94.62 ± 0.10	63.6 ± 2.55

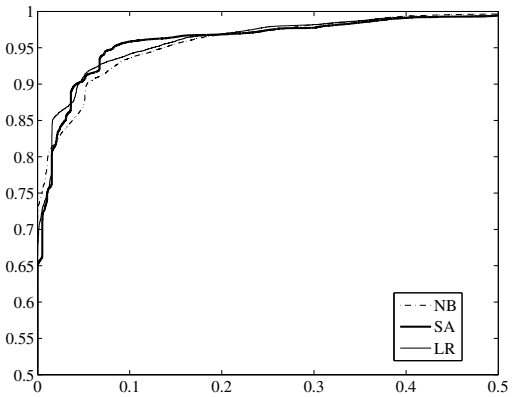
ROC-кривые в осях X:(1−специфичность), Y:чувствительность



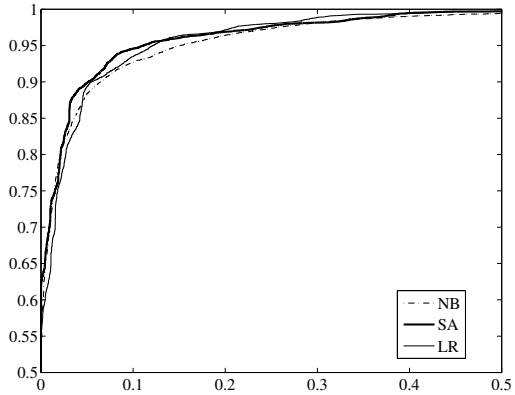
асептический некроз ГБК



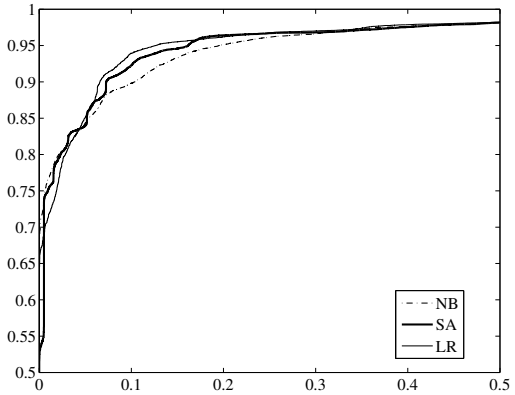
желчнокаменная болезнь



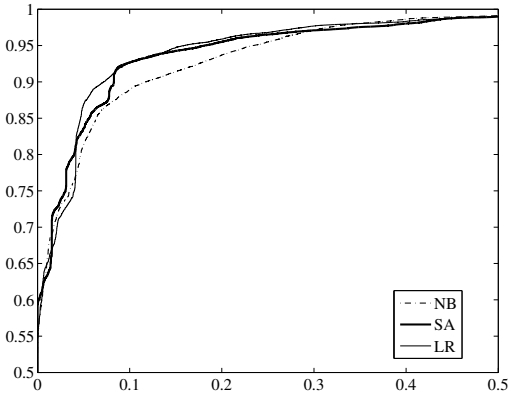
ишемическая болезнь



хронический гастрит 1



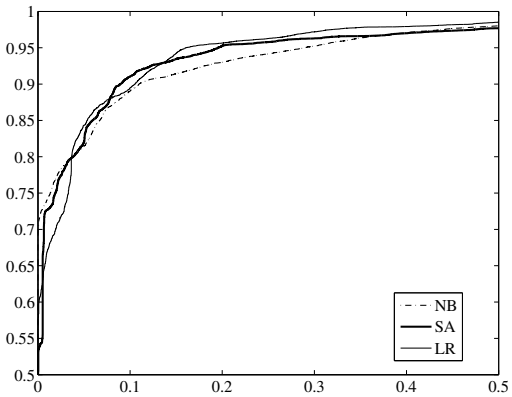
сахарный диабет



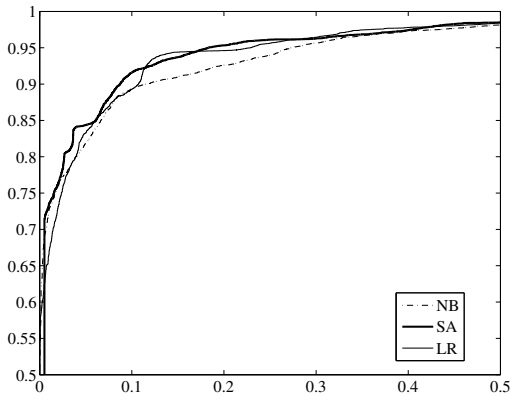
гипертония

NB — Naïve Bayes, SA — Syndrome Algorithm, LR — Logistic Regression

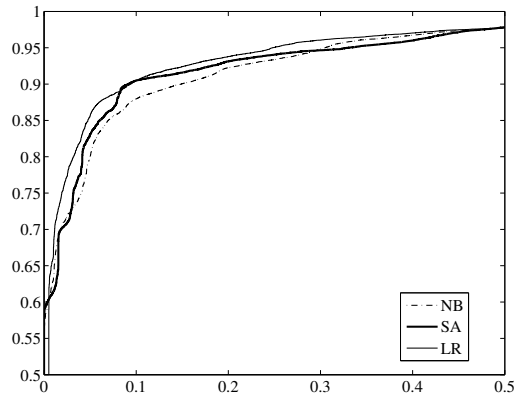
ROC-кривые в осях X:(1−специфичность), Y:чувствительность



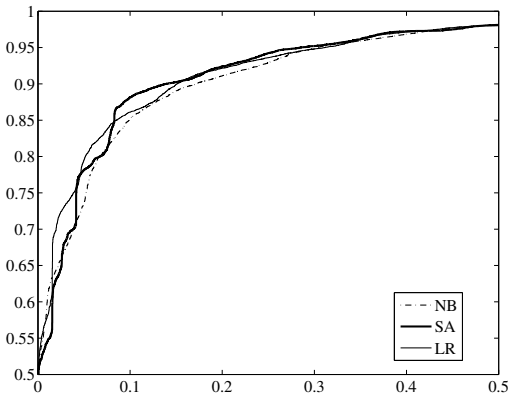
рак общий



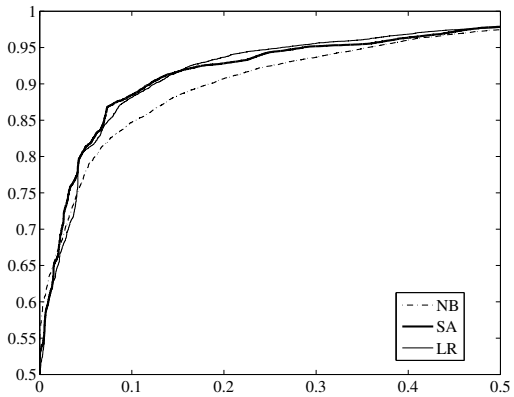
аденома простаты



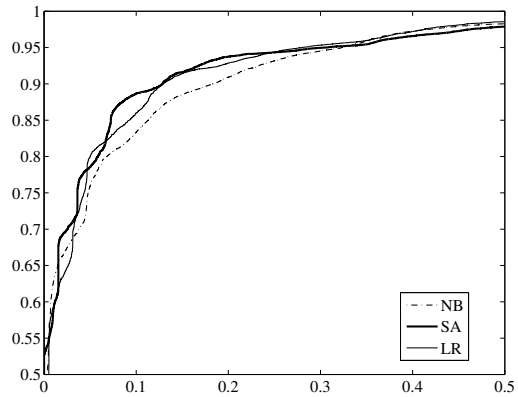
зоб щитовидной железы



хронический гастрит 2



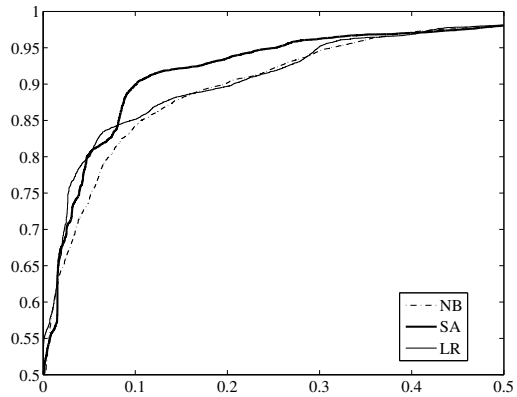
дискинезия ЖВП



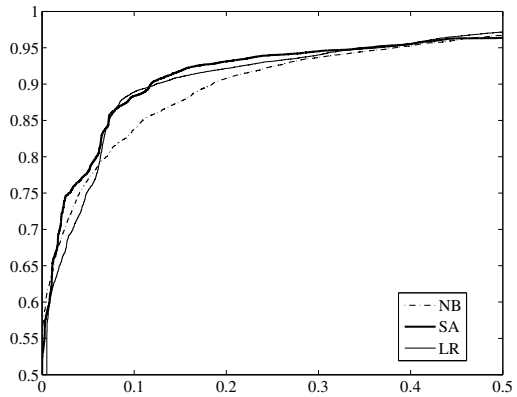
мочекаменная болезнь

NB — Naïve Bayes, SA — Syndrome Algorithm, LR — Logistic Regression

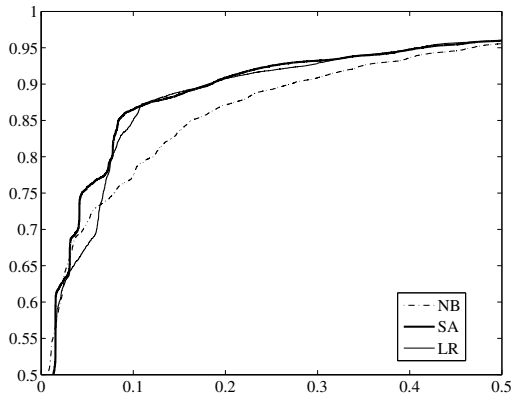
ROC-кривые в осях X:(1−специфичность), Y:чувствительность



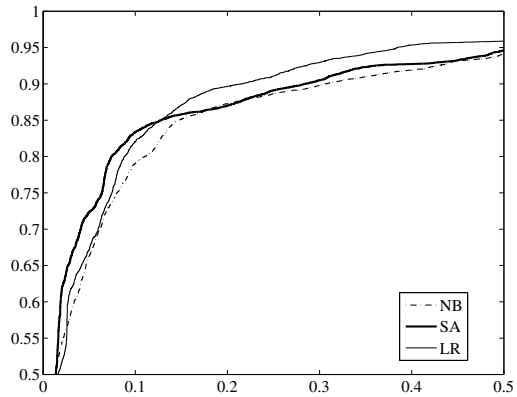
хронический холецистит



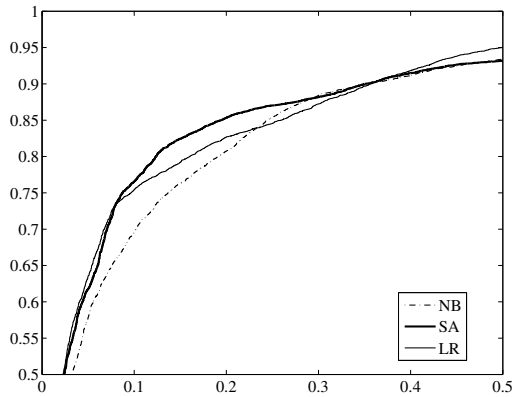
язвенная болезнь



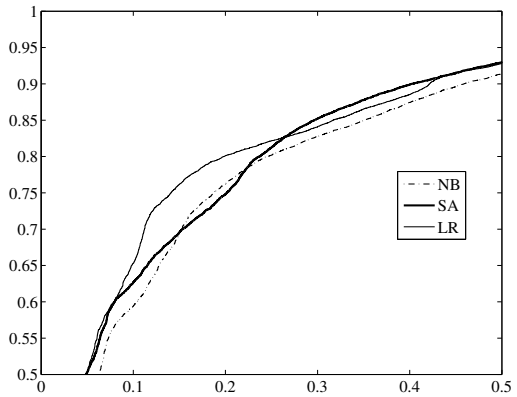
миома матки



хронический аднексит



анемия

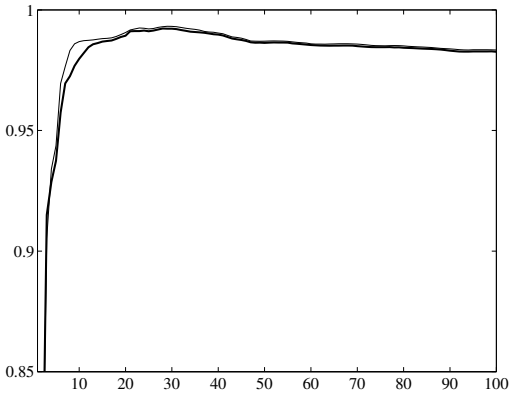


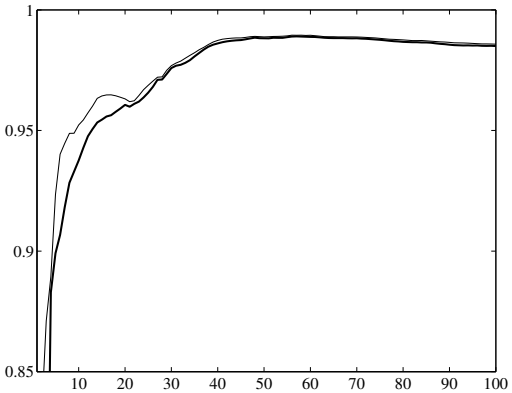
вегетососудистая дистония

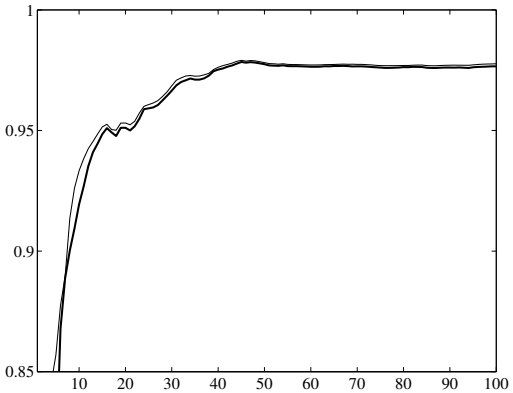
NB — Naïve Bayes, SA — Syndrome Algorithm, LR — Logistic Regression

Зависимости AUC от числа используемых признаков K

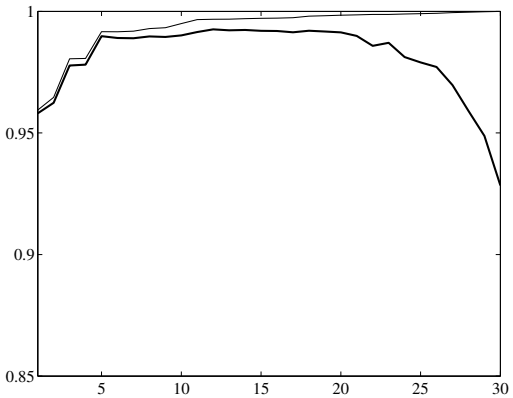
Синдромный алгоритм (наивный Байес на K признаках):

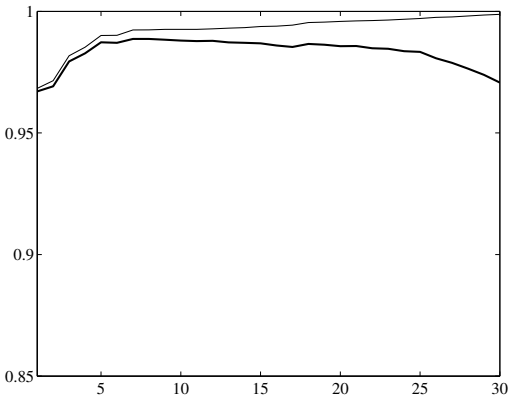


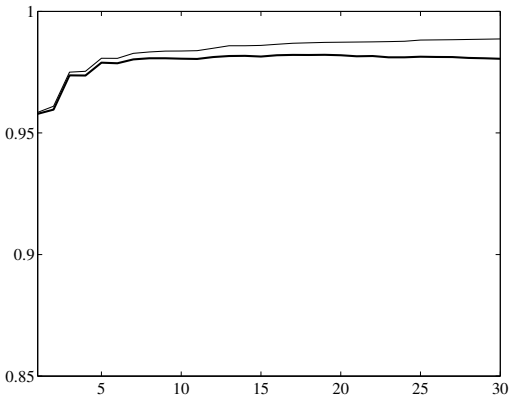




Логистическая регрессия (K — число главных компонент):





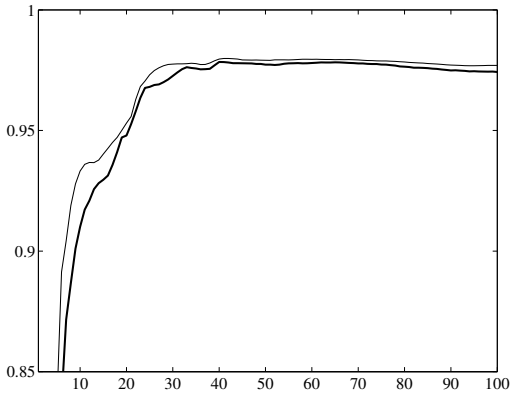


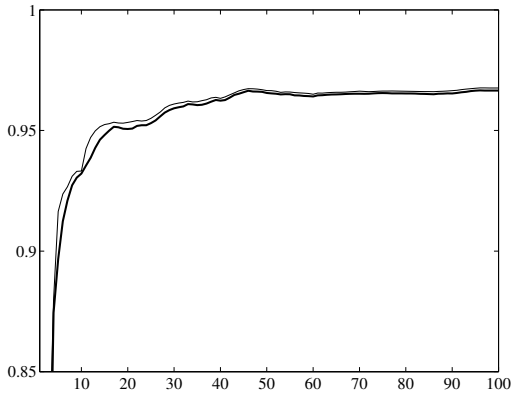
асептический некроз ГБК | желчнокаменная болезнь | ишемическая болезнь

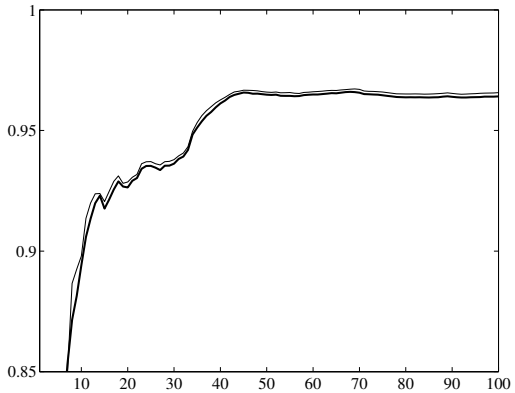
Тонкая (верхняя) линия — на обучающей выборке
Толстая (нижняя) линия — на тестовой выборке

Зависимости AUC от числа используемых признаков K

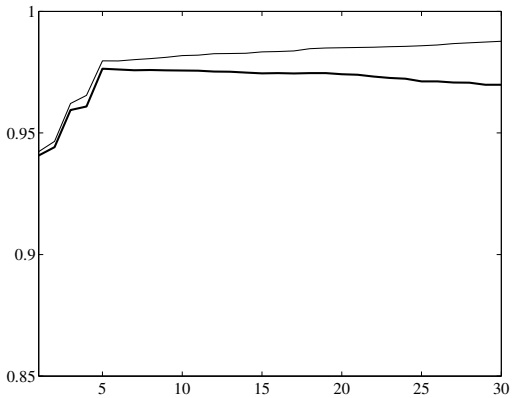
Синдромный алгоритм (K — число признаков):

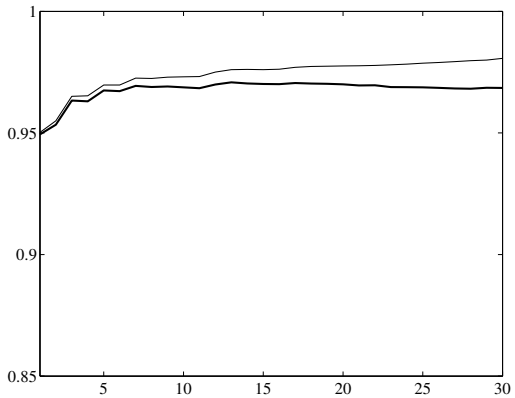


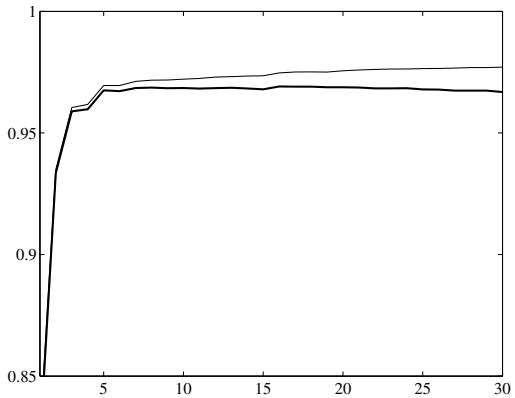




Логистическая регрессия (K — число главных компонент):





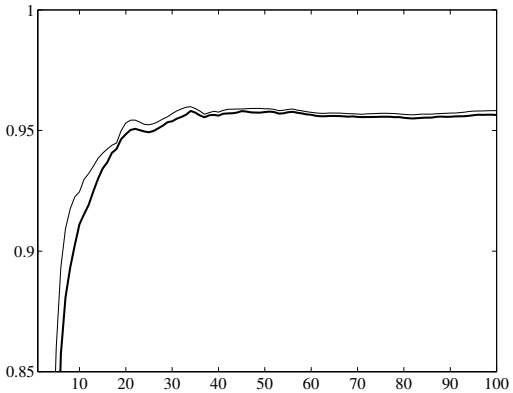


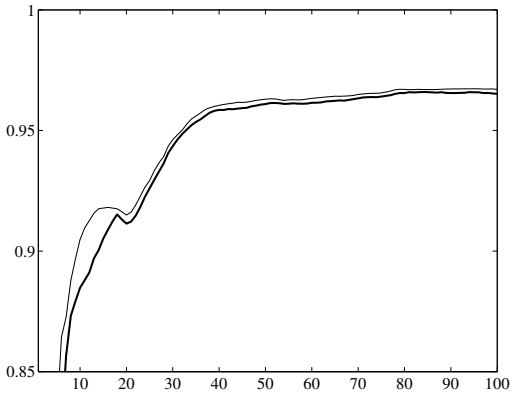
хронический гастрит 1 | сахарный диабет | гипертония

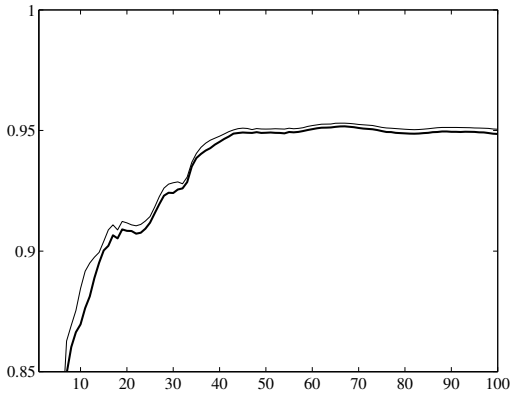
Тонкая (верхняя) линия — на обучающей выборке
Толстая (нижняя) линия — на тестовой выборке

Зависимости AUC от числа используемых признаков K

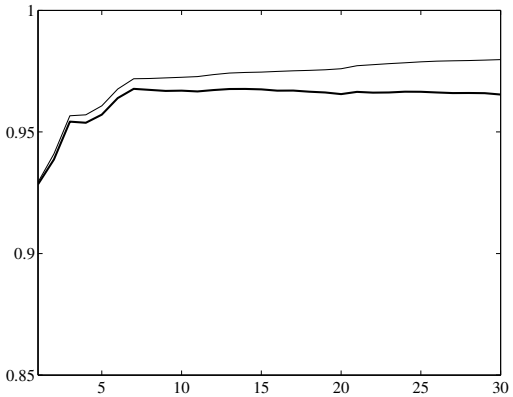
Синдромный алгоритм (K — число признаков):

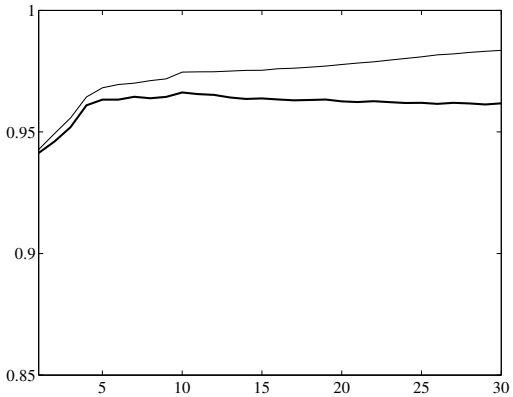


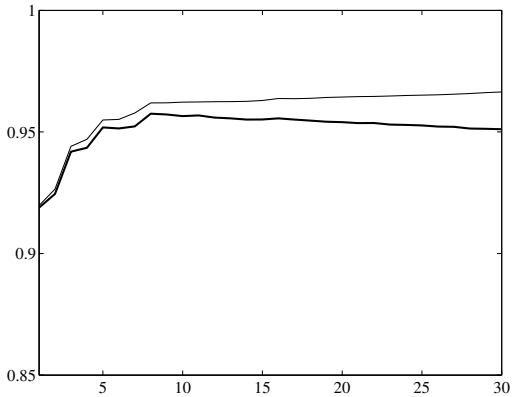




Логистическая регрессия (K — число главных компонент):





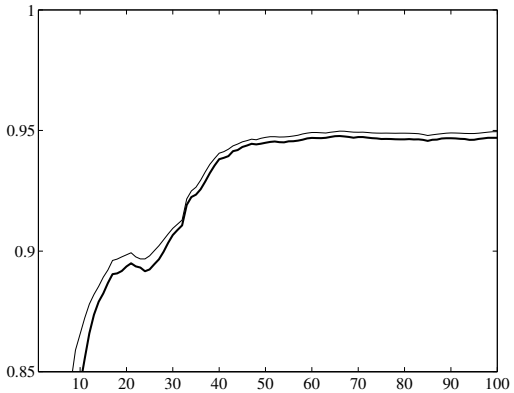


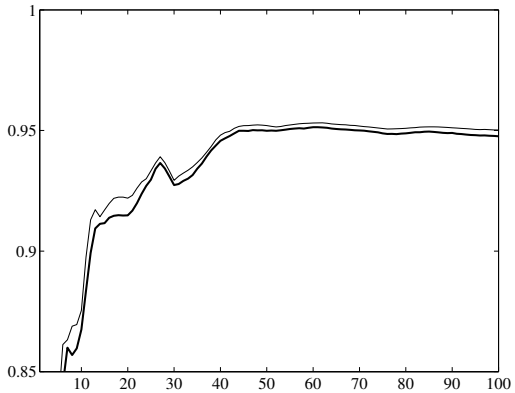
рак общий | аденома простаты | зоб щитовидной железы

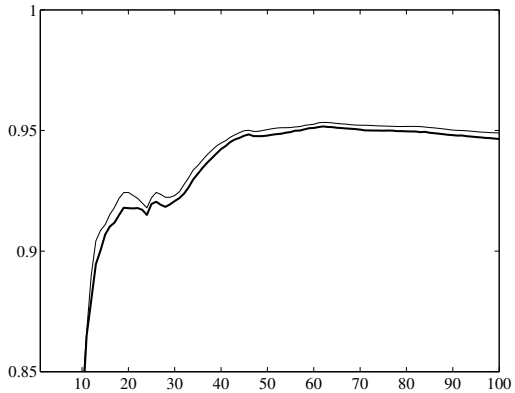
Тонкая (верхняя) линия — на обучающей выборке
Толстая (нижняя) линия — на тестовой выборке

Зависимости AUC от числа используемых признаков K

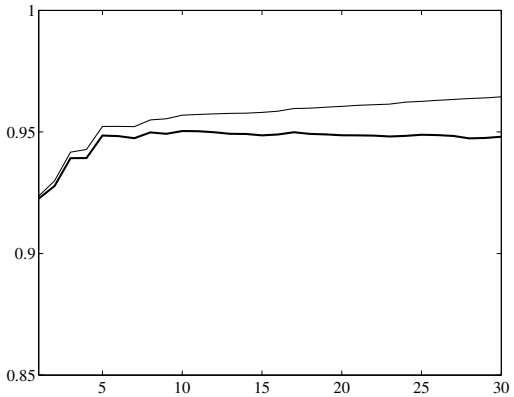
Синдромный алгоритм (K — число признаков):

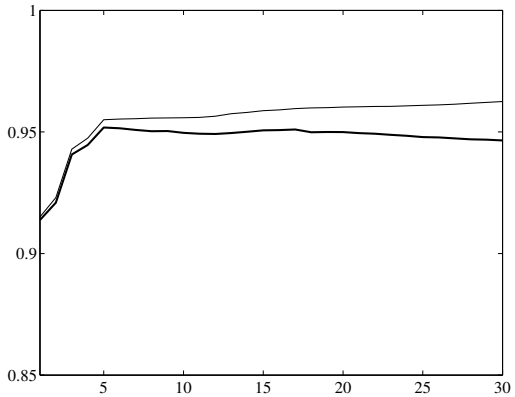


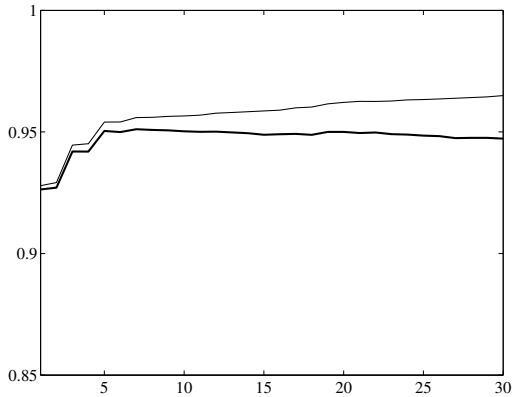




Логистическая регрессия (K — число главных компонент):





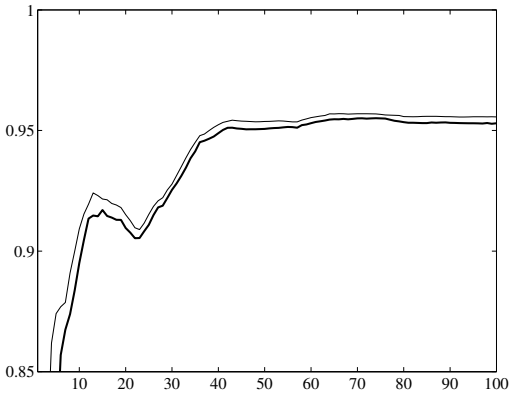


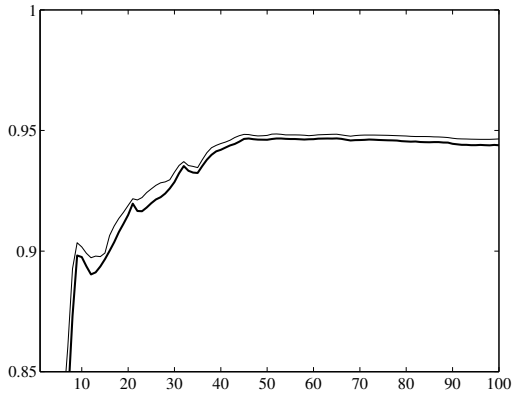
хронический гастрит 2 дискинезия ЖВП мочекаменная болезнь

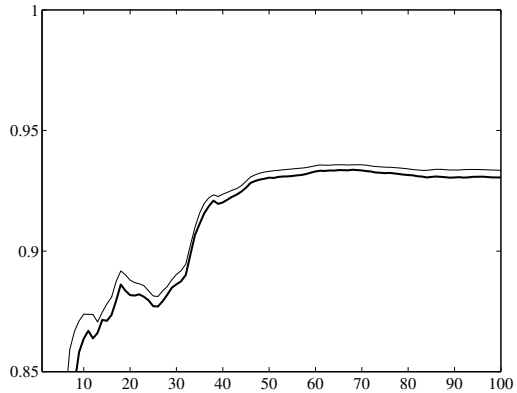
Тонкая (верхняя) линия — на обучающей выборке
Толстая (нижняя) линия — на тестовой выборке

Зависимости AUC от числа используемых признаков K

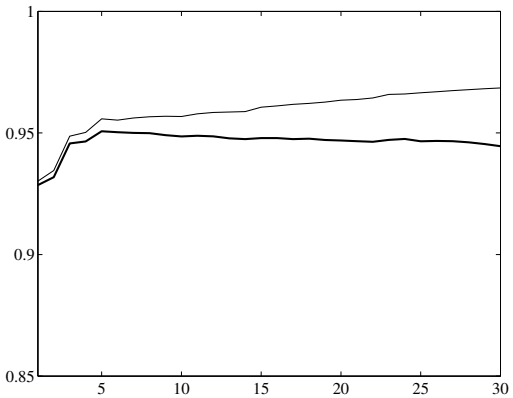
Синдромный алгоритм (K — число признаков):

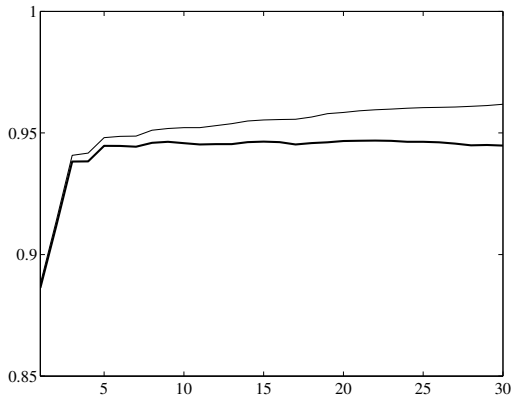


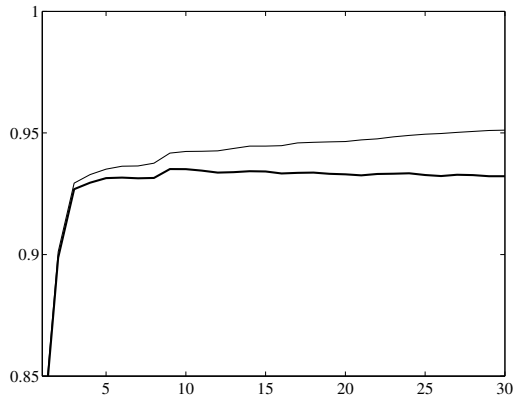




Логистическая регрессия (K — число главных компонент):





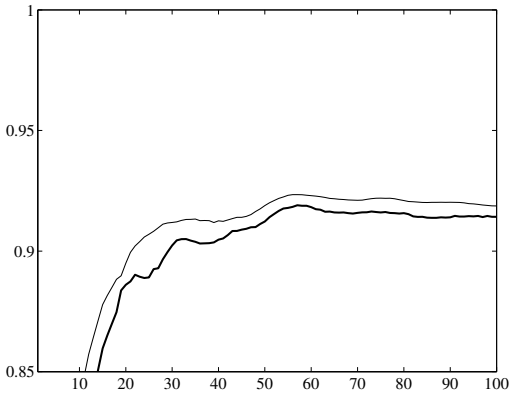
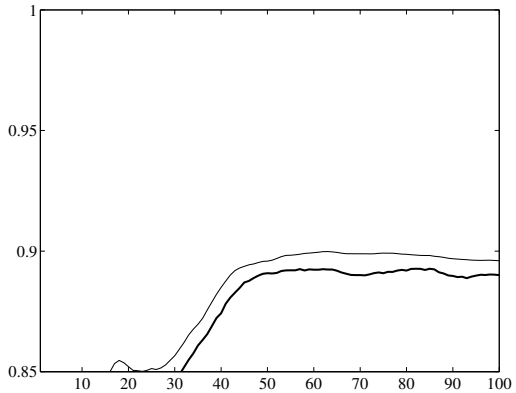
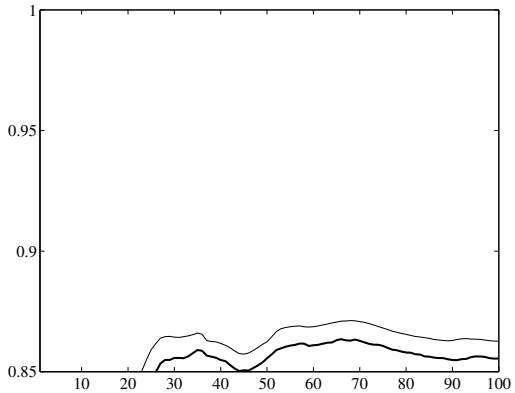


хронический холецистит | язвенная болезнь | миома матки

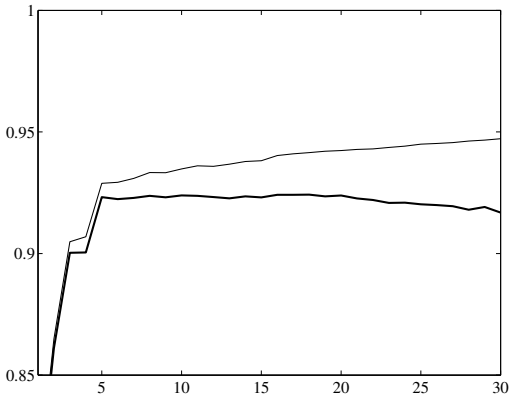
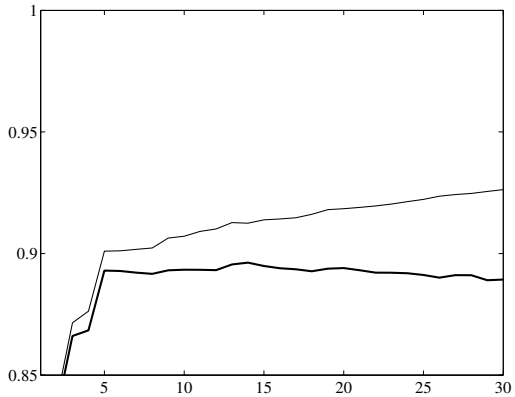
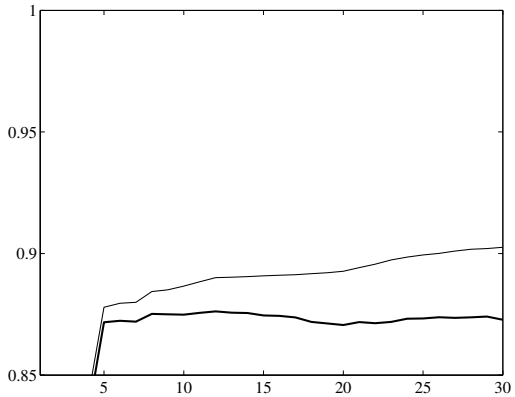
Тонкая (верхняя) линия — на обучающей выборке
Толстая (нижняя) линия — на тестовой выборке

Зависимости AUC от числа используемых признаков K

Синдромный алгоритм (K — число признаков):

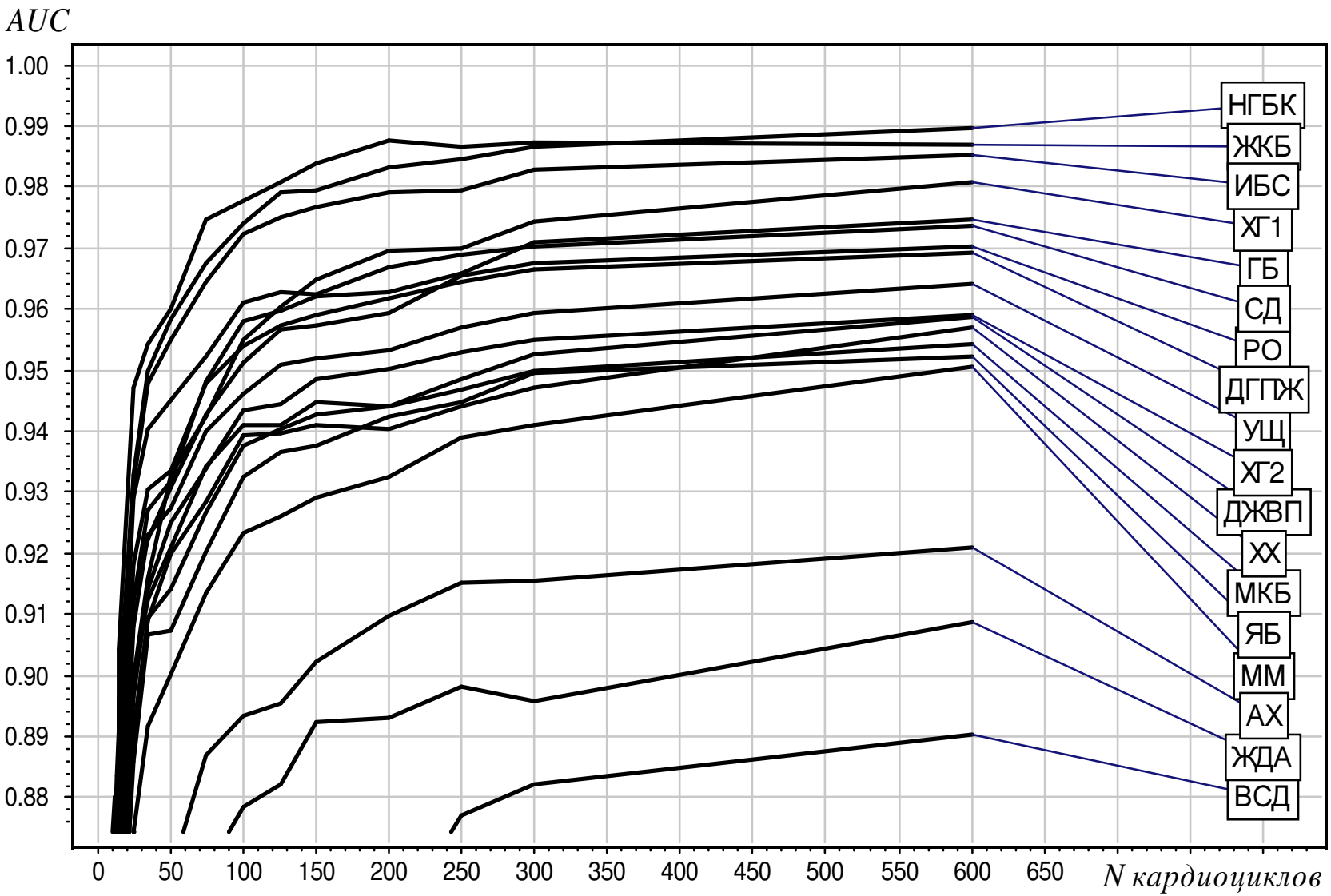
Логистическая регрессия (K — число главных компонент):

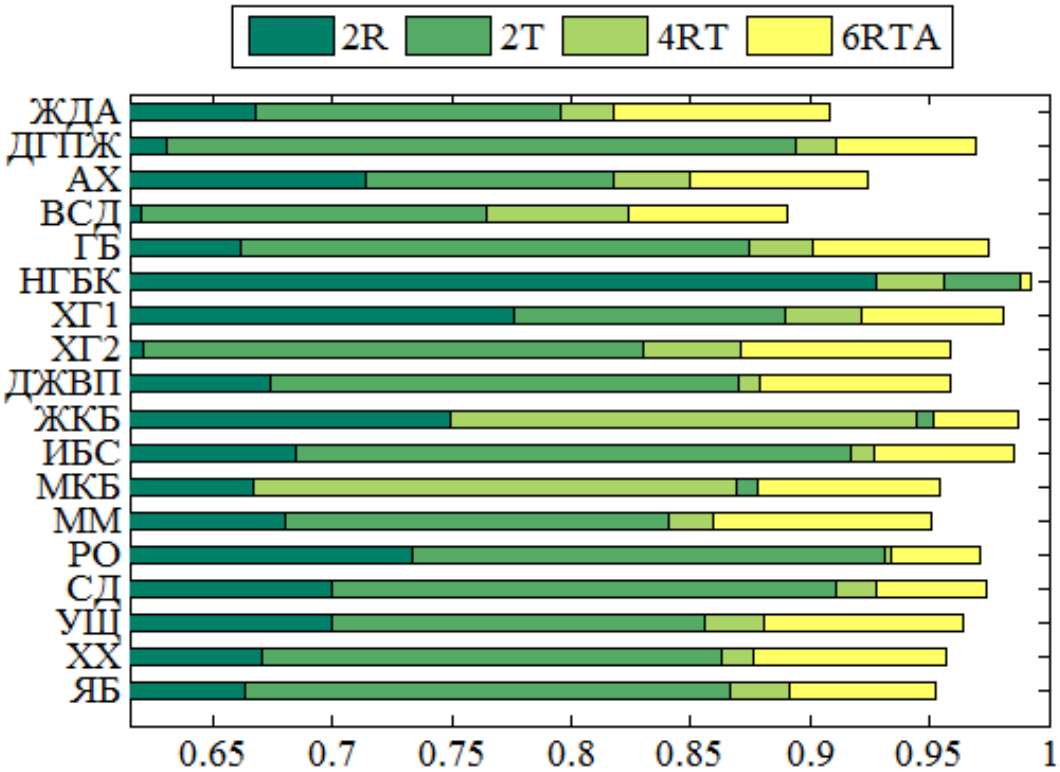
хронический аднексит | анемия | вегетососудистая дистония

Тонкая (верхняя) линия — на обучающей выборке
Толстая (нижняя) линия — на тестовой выборке

Зависимость AUC от длительности регистрации ЭКГ



Зависимость AUC от типа символьного кодирования



2R: 2-символьная, только приращения амплитуд

2T: 2-символьная, только приращения интервалов

4RT: 4-символьная, приращения интервалов и амплитуд

6RTA: 6-символьная, приращения интервалов, амплитуд и их отношений

Открытые данные по инфарктам миокарда: база данных PTB

Данные национального метрологического института Германии.

Число записей ЭКГ-сигналов: 320 больных, 74 здоровых.
Длительность регистрации ЭКГ: 100–200 кардиоциклов.

AUC при 6-символьном кодировании (6RTA) для трёх методов:
LR — логистическая регрессия,
RF — случайный лес,
SA — наивный Байес с отбором признаков

	LR	RF	SA
2-граммы	87.7	87.9	86.1
3-граммы	89.4	**89.6**	87.1
4-граммы	88.6	87.7	86.9

Bousseljot R., Kreiseler D., Schnabel A. Nutzung der EKG-Signaldatenbank CARDIODAT der PTB über das Internet. Biomedizinische Technik. 1995.

Выводы

- Удивительно высокая точность диагностики
  — качество диагностики подтверждено кросс-валидацией
  — в том числе в экспериментах на открытых данных
- Статистически обоснованы основные элементы технологии информационного анализа ЭКГ-сигналов:
  — выбор диагностических эталонов болезней
  — выбор длительности регистрации сигнала (300–600)
  — выбор оптимального алфавита и типа кодирования
  — выбор методов машинного обучения

Воронцов Константин Вячеславович
voron@forecsys.ru

www.MachineLearning.ru • Участник:Vokov