

# **Построение модели времени дожития пациентов с боковым амиотрофическим склерозом на основе графов водородных связей**

Воронкина Д.К, ФПМИ НГТУ.  
Научный руководитель: Тимофеев В.С д.т.н

# Структура данных

	A4T	A4V
<u>4.A@N-20.A@O</u>	0.821	0.851
<u>5.A@N-150.A@O</u>	0.906	0.974
<u>6.A@N-18.A@O</u>	0.928	0.899

Столбец с информацией о ребрах: донор-акцептор

Донор

Акцептор

Столбцы, содержащие веса соответствующих ребер. Каждый столбец представляет информацию по одному пациенту.



# Модель регрессии

Рассматриваемое уравнение:

$$Y = XB + E, \text{ где}$$

$X$  – матрица объясняющих переменных, составленная из вычисленных характеристик графа;

$Y$  – вектор наблюдаемых значений, составленный из времени дожития пациентов;

$B$  – неизвестные коэффициенты регрессии;

$E$  – случайная ошибка регрессии.

# Интерпретация данных

6.A@N

Позиция  
аминокислотного  
остатка

Имя субъединицы

Название атома

4.A@N

20.A@0

1. В качестве вершины выступает водородная связь полностью.

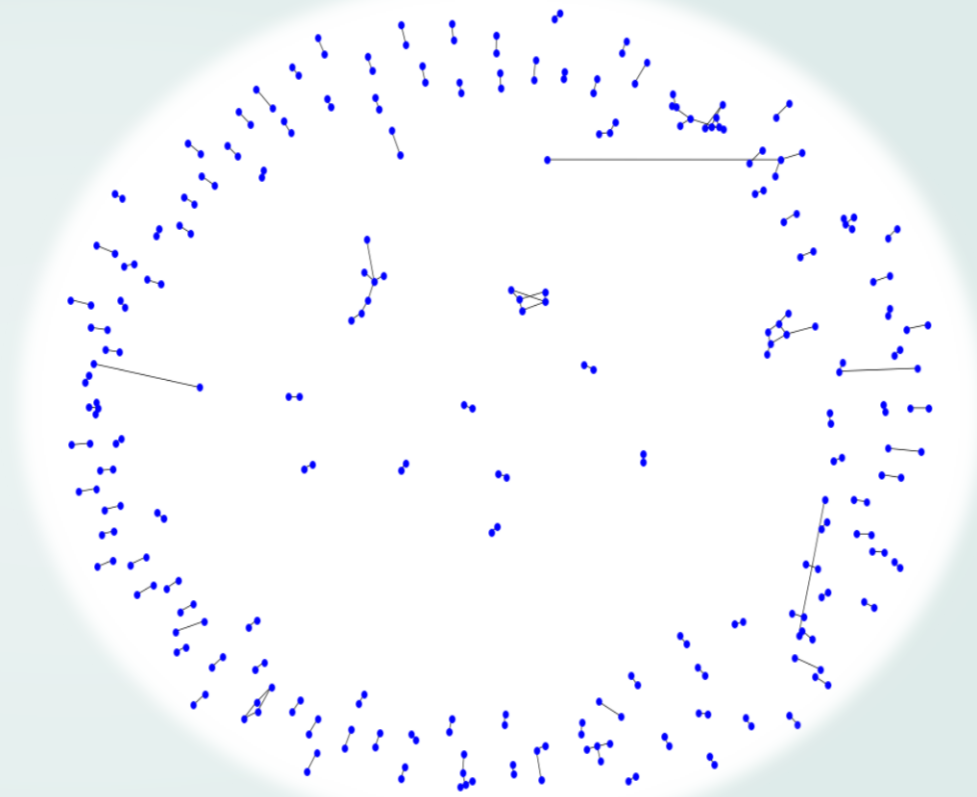


Рис. 1 – Пример графа для одного из пациентов.

20.A

4.A

2. В качестве вершины выступает позиция аминокислотного остатка с субъединицей.

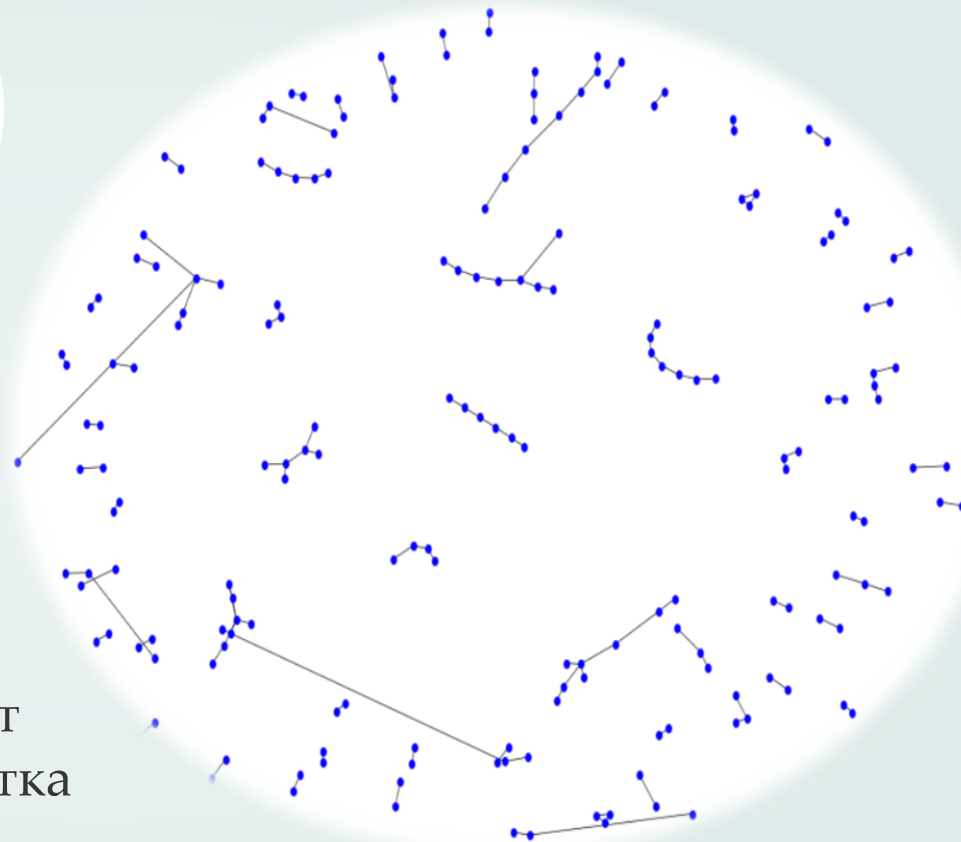




Рис. 2 – Пример графа для одного из пациентов.



# Характеристики, используемые для графа.

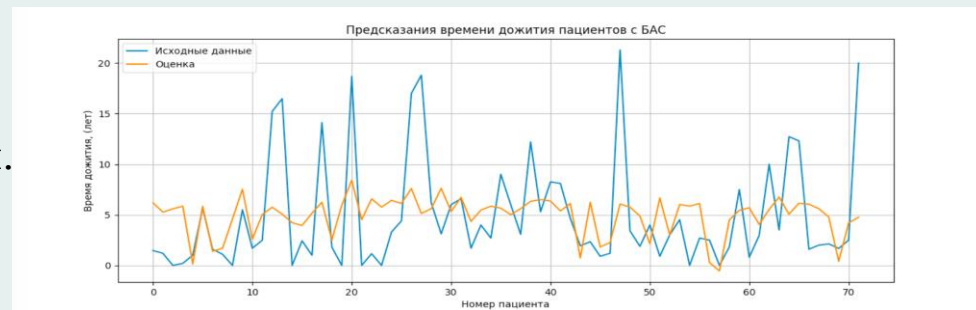


- Коэффициент кластеризации;
- Модульность;
- Определитель матрицы смежности с весами;
- Определитель матрицы Лапласа;
- Ребро с наименьшим весом;
- Сумма всех ребер графа.
- Количество вершин;
- Количество ребер;
- Плотность графа;
- Радиус графа;
- Диаметр графа;
- Транзитивность;
- Средний коэффициент кластеризации;
- Связность графа;
- Степень ассортивности графа;
- Индекс Эстрады;
- Размер наибольшего клика на графе;



Рис. 4 - Модель для графа, где в качестве вершины выступает полная водородная связь. Характеристики посчитаны для всего графа.  
Коэффициент детерминации 0.139

Рис. 5 - Модель для графа, где в качестве вершины выступает позиция аминокислотного остатка с субъединицей. Характеристики посчитаны для всего графа.  
Коэффициент детерминации 0.121





20.A@0

4.A@N

В качестве вершины выступает водородная связь полностью.

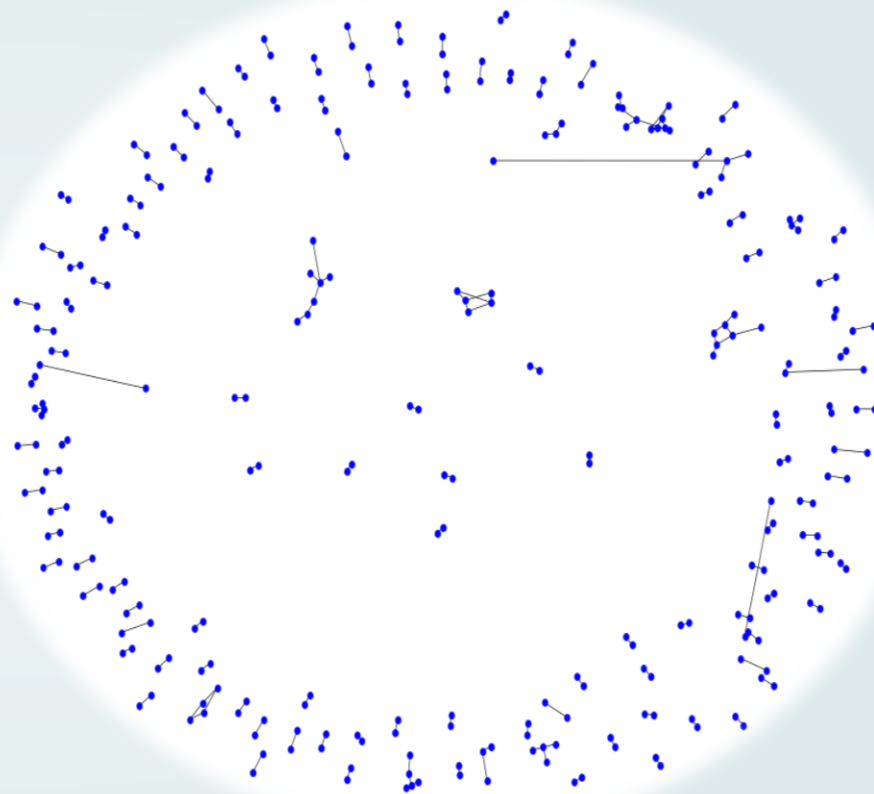


Рис. 7 – Пример графа для одного из пациентов.

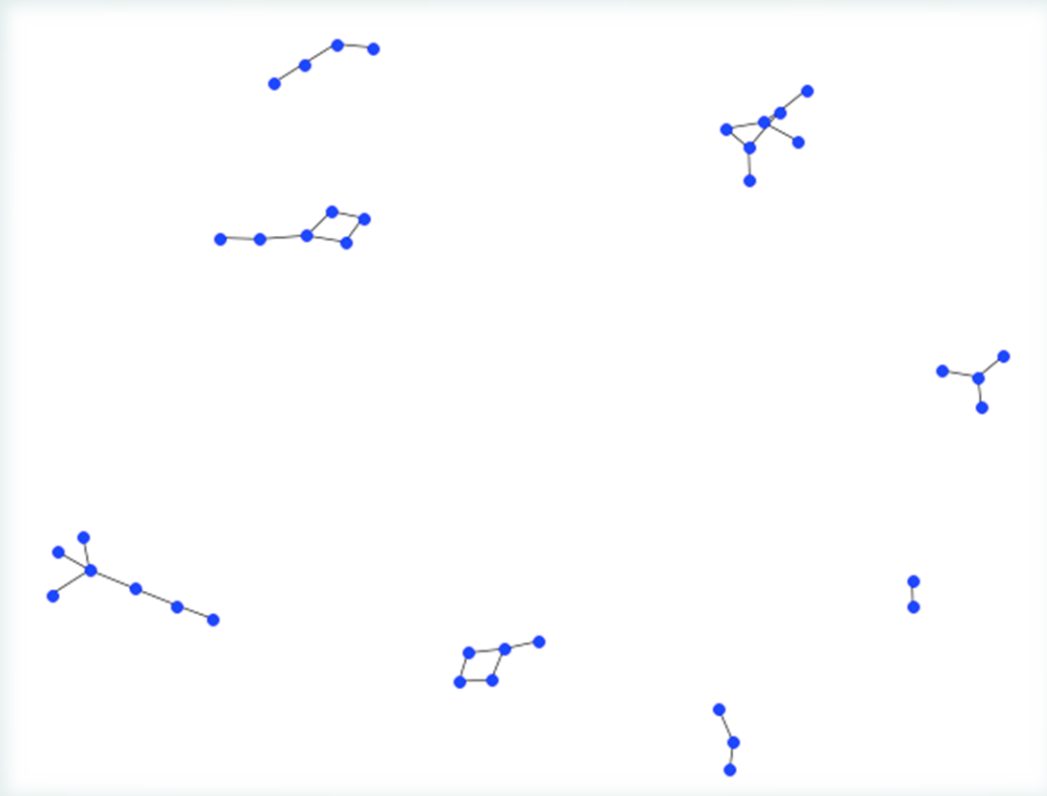


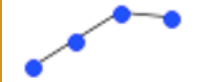




Рис. 8 – Подграфы с уникальной структурой для одного из пациентов

# Пример возможной структуры графа

			
Пациент №1	+	-	+ (но с другими вершинами)
Пациент №2	-	+ (но с другими вершинами)	+
Пациент №3	+ (но с другими вершинами)	-	-



# Характеристики, используемые для графа.



- Коэффициент кластеризации;
- **Модульность;**
- Определитель матрицы смежности с весами;
- Определитель матрицы Лапласа;
- **Ребро с наименьшим весом;**
- **Сумма всех ребер графа.**
- Количество вершин;
- Количество ребер;
- Плотность графа;
- Радиус графа;
- Диаметр графа;
- Транзитивность;
- Средний коэффициент кластеризации;
- Связность графа;
- Степень ассортивности графа;
- Индекс Эстрады;
- Размер наибольшего клика на графе;

# Результаты

Используемая характеристика	Коэффициент детерминации	Кол-во регрессоров	Кол-во значимых регрессоров
Модульность	0,337	35	5
Ребро с наименьшим весов	0,502	34	9
Сумма всех путей	0,219	37	3

# Результаты

	Коэффициент детерминации	Кол-во регрессоров	Кол-во значимых регрессоров
Общая выборка МНК	0,515	15	9

## Характеристики, которые вошли в модель:

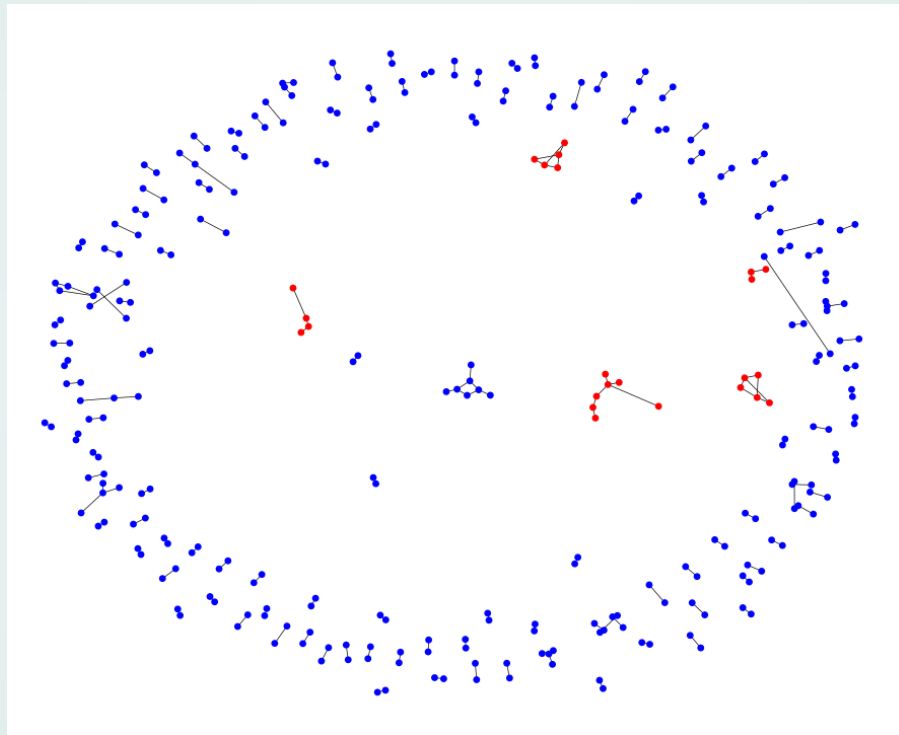
Короткий путь для подграфов:

- 101.A@OD1\_79.A@NH2\_101.A@OD2\_79.A@NH1\_81.A@O\_74.A@O\_83.A@O
- 124.B@OD1\_46.B@NE2\_71.B@NE2\_124.B@OD2\_126.B@N
- 124.B@OD1\_46.B@NE2\_71.B@NE2\_124.B@OD2\_126.B@N\_an
- 54.B@OG1\_52.B@OD1\_52.B@OD2\_54.B@N
- 126.A@N\_124.A@OD2\_124.A@OD1\_71.A@NE2\_46.A@NE2
- 46.B@NH2\_124.B@OD2\_124.B@OD1\_71.B@NE2\_126.B@N
- 125.B@OD2\_71.B@NE2\_124.B@OD1\_126.B@N\_124.B@OD2\_46.B@NE2
- 124.B@OD1\_46.B@NE2\_71.B@NE2\_126.B@N\_124.B@OD2

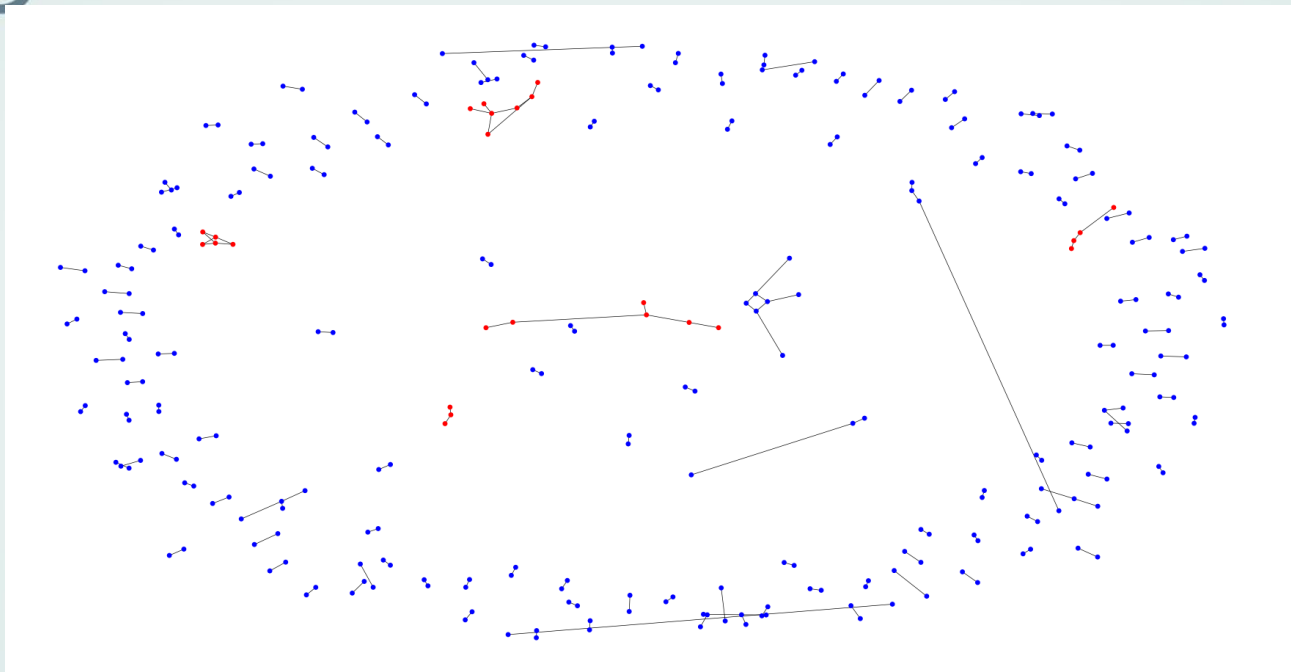
Сумма всех ребер для подграфов:

- 79.A@NE\_80.A@O\_83.A@N

## Выделенные значимые подграфы для пациента №1

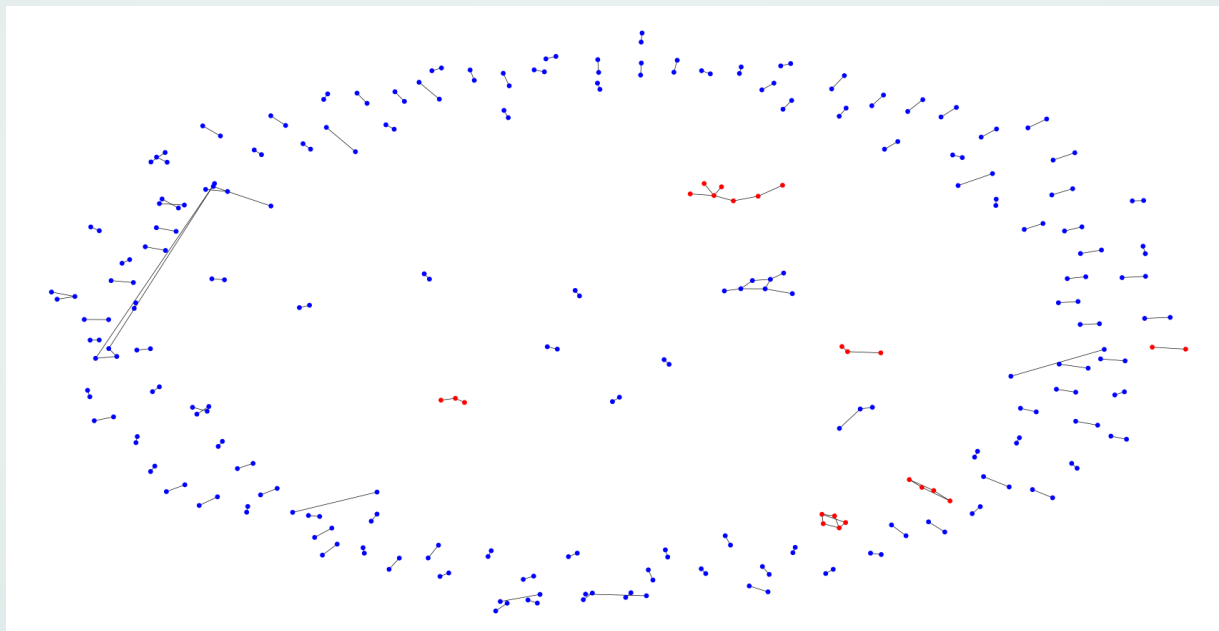


## Выделенные значимые подграфы для пациента №2





## Выделенные значимые подграфы для пациента №3



# Оценки

	Обычный МНК	Робастный Хьюбер	Random Forest
RMSE	3.78	3.86	2.09
R2	0.515	-	0.86

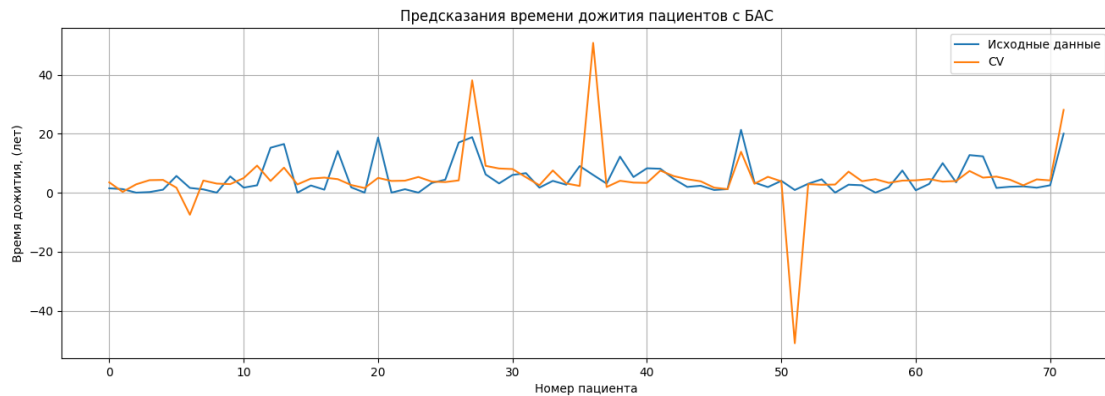
# Кросс-валидация

	п отсутствующих значений	Обычный МНК	Робастный Хьюбер	Random Forest
Ошибка кросс валидации	1	11.4812028	11.78645428	5.9304524865
	6	11.91748789	13.06804192	6.2375873517
	9	10.03913134	11.07664542	6.1287852839
	12	11.23820801	12.38048862	6.1448774647

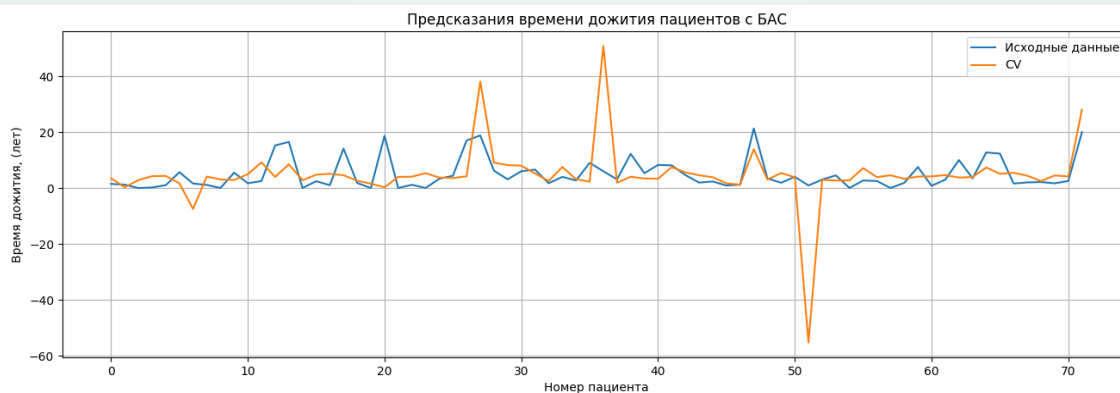
$$CV = \sum_{i=1}^N (y_i - \hat{y}_{(-i)})^2$$

где,  $\hat{y}_{(-i)}$  - полученные значения после оценки с учетом того, что отсутствует i-ая строка.

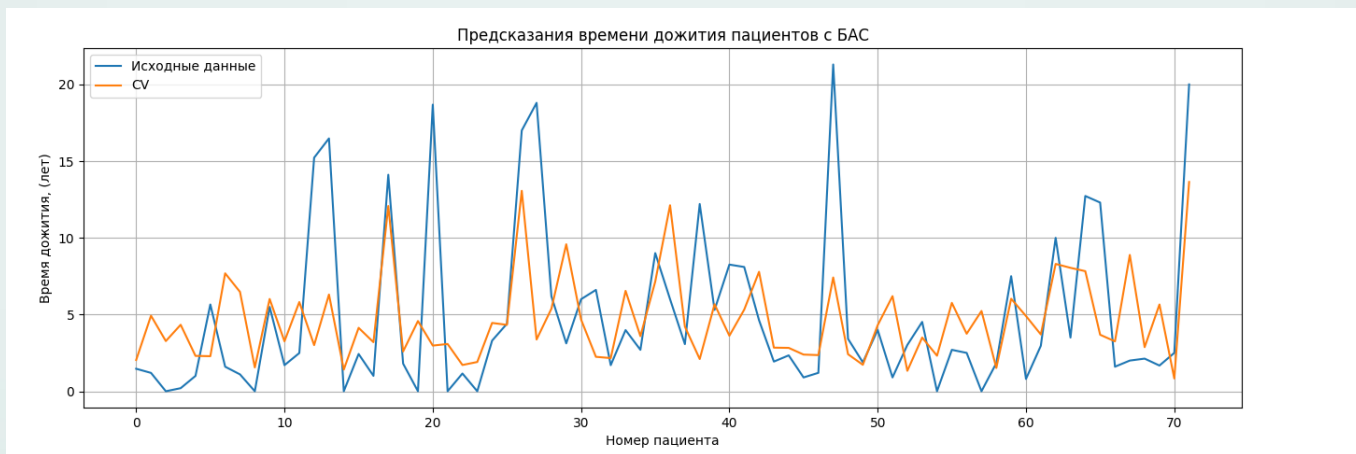
# Ordinary Least Squares



# Robust Method




# Random Forest



## Коэффициенты регрессии

	Ordinary	Robust
x1	1.17	0.85
x2	-14.49	-13.56
x3	10.2	9.61
x4	1.76	1.89
x5	7.91	7.53
x6	1.6	1.71
x7	1.71	1.79
x8	6.08	5.8
x9	-1.56	-1.32
const	4.93	4.28



# **Построение модели времени дожития пациентов с боковым амиотрофическим склерозом на основе графов водородных связей**

Воронкина Д.К, ФПМИ НГТУ.  
Научный руководитель: Тимофеев В.С д.т.н