

個人の遺伝子情報(SNP)と生活習慣病の関連性  
を用いた健康診断支援システム

平成 24 年度修了

三重大学大学院 地域イノベーション学研究科  
博士前期課程 地域イノベーション学専攻  
先端融合工学ユニット

森上 斐斗

# 目次

Abstract .....	1
第1章 はじめに.....	2
第2章 遺伝子のSNP .....	4
2.1 塩基配列 .....	4
2.2 SNP (スニップ) .....	5
2.3 SNPの違いによる体質の変化.....	6
第3章 既存の健康診断支援システム .....	7
3.1 健康診断システムの必要性.....	7
3.2 既存の健康診断支援システム .....	7
3.1 Genomarker での検査方法.....	8
3.2 Genomarker の検査結果 .....	9
第4章 データの分析 .....	11
4.1 遺伝子・生体指標集積データ .....	11
4.2 データの分布.....	13
第5章 生活習慣病を予測するシステム .....	15
5.1 提案システム.....	15
5.2 的確な生活指導を実現する出力.....	16
第6章 提案システムの概要 .....	17
6.1 学習過程 .....	18
6.1.1 特定疾患の選択 .....	18

6.1.2	データの基準化 .....	19
6.1.3	診断用辞書の作成.....	21
6.2	健康診断過程.....	24
6.2.1	診断用辞書の選択.....	24
6.2.1	発症危険度の算出.....	25
6.2.2	発症危険性推定曲線の作成.....	27
<b>第7章</b>	<b>評価実験.....</b>	<b>29</b>
7.1	実験1 —SNPによる発症危険度の違い—.....	29
7.1.1	実験方法 .....	29
7.1.2	実験結果 .....	31
7.2	実験2—個人の発症危険度推定曲線—.....	32
7.2.1	実験方法 .....	32
7.2.2	実験の結果.....	33
7.3	実験の考察 .....	35
<b>第8章</b>	<b>終わりに.....</b>	<b>36</b>
<b>謝辞</b>	<b>.....</b>	<b>37</b>
<b>参考文献</b>	<b>.....</b>	<b>38</b>
<b>発表論文リスト</b>	<b>.....</b>	<b>39</b>

## 図 一覧

1-1	日本人の死因割合（10万人対）（2011） .....	3
2-1	DNAの構造（塩基対） .....	4
2-2	塩基配列の一般的表現.....	4
2-3	SNP .....	5
2-4	SNPによる体質の違い（一例） .....	6
3-1	糖尿病の診断結果（例） .....	10
4-1	発症者の分布状況.....	14
4-2	あるSNPの分布状況.....	14
5-1	提案システムの構想 .....	15
5-2	提案システムの出力（発症危険度推定曲線） .....	16
6-1	研究の流れ.....	17
6-2	辞書作成用データの分類.....	21
6-3	特徴ベクトルの分類（モデル） .....	22
6-4	特徴平均ベクトルの作成（モデル） .....	23
6-5	疾患別・SNP別診断用辞書.....	23
6-6	被験者の診断用辞書を選択 .....	24

6-7	入力ベクトルとの比較 (モデル) .....	25
6-8	発症危険度推定曲線の作成 .....	28
7-1	実験結果 .....	31
7-2	発症者 50 名の発症危険度推定曲線 .....	33
7-3	非発症者 50 名の発症危険度推定曲線 .....	34

## 表 一覧

3-1 Genomarker で検査可能な項目 .....	9
4-1 遺伝子・生体指標集積データに記載されている情報 .....	12
6-1 SNPの変換ルール（例） .....	19
6-2 基準化後の値（例） .....	20
7-1 統一した値.....	30

# Health Check Support System Using Relationship between Individual Genome Information (SNP) and Lifestyle Disease

Ayato Morikami

March 2013

## 1. Introduction

A genetic analysis is developing to identify single nucleotide polymorphism (SNP) associated with lifestyle disease. Many researchers are investigating the lifestyle disease depended on one SNP. However, almost onset influence is not clear for the combination of patient's SNPs.

This object of this research is the proposal of a new risk estimation method of lifestyle disease in the consideration of the relevance of SNP and biomarkers. By using this system, a physician can suggest lifestyle guidance more exactly to a subject.

## 2. Proposed Method

### 2.1. Outline of System

Figure.1 shows the flowchart of my proposed system.

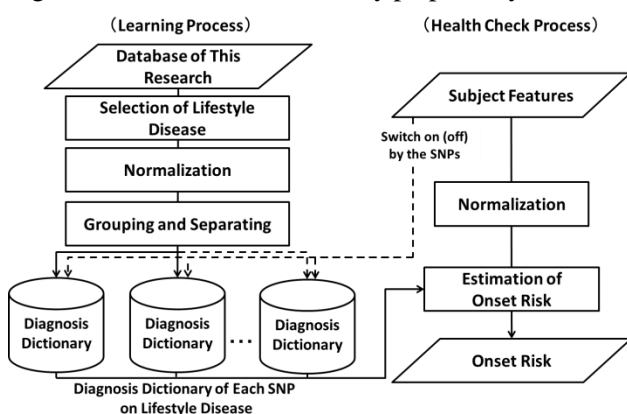


Figure 1 Flowchart of proposed system

### 2.2. Learning Process

In learning process, the system generates diagnosis dictionaries from many sets of SNPs and biomarkers such as BMI, age, sex, smoking, blood pressure, cholesterol, HbA1c. The sets are normalized for each features scale.

Diagnosis dictionaries are generated for each SNP types on specific disease.

### 2.3. Health Check Process

To estimate onset risk, subject feature is compared to individual diagnosis dictionaries. Individual diagnosis dictionaries are selected by subject's SNP types.

## 3. Evaluation of Proposed System

### 3.1. Materials for experiment

Specific disease is myocardial infarction, and variable element is systolic blood pressure. Numbers of onset samples are 1461 persons and non-onset samples are 649 persons to generate diagnosis dictionaries. The number of samples is 6 persons and these biomarker values are made even.

### 3.2. Experimental Result

Figure 2 shows that the onset risk of acute myocardial infarction is different for each subject (SNPs type) under the same systolic blood pressure.

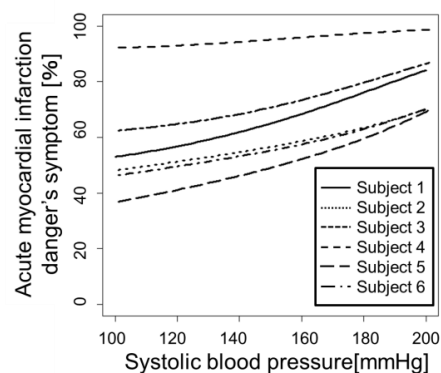


Figure 2 Experimental Result

I am considering that the result would become more appropriate for medical doctors.

## 4. Conclusions

We proposed a new health check system using SNPs and biomarkers. We will consider the other disease.

## References

- [1] GenoMarker, G&GScienceCo, <http://www.genomarker.jp/index.html>
- [2] Yoshiji Yamada, Hitoshi Matsuo, Sachiro Watanabe, Kimihiko Kato, Takeshi Hibino, Kiyoshi Yokoi, Sahoko Ichihara, Norifumi Metoki, Hidemi Yoshida, Kei Satoh, Yoshinori Nozawa. : Association of a polymorphism of CYP3A4 with type 2 diabetes mellitus. *Int J Mol Med*, Vol.20, No.5, pp.703-707, (2007)

# 第1章

## はじめに

現在, 日本人の6割以上はがん, 心臓病, 脳卒中などの生活習慣病で死亡している (図1-1) [1]. 生活習慣病は日常の生活習慣の乱れの積み重ねが大きな要因となって発症し, 初期段階では自覚症状がほとんどない. そこで, 学校や職場等では定期的な健康診断を行っており, 病気が出す信号(異常値)を検出することで, 生活習慣病の早期発見と早期治療に効果を上げている [2].

また近年, 遺伝子解析技術とコンピュータの発達により, 高速遺伝子シーケンサが開発され, 安価で高速な遺伝子解析が可能となりつつある. これに伴い, 遺伝子を分析することで様々な生体情報を得る研究が活発に行われている.

遺伝子には人体のあらゆる情報が詰まっている. その情報の中には病気のかかりやすさや薬の効き方などの個人の体質も含まれている. そのため, 遺伝子を分析することにより個人の体質を明らかにすることが可能となると考えられている. 医療の分野では, この遺伝子から得られる個人の体質をもとに, 最適な投薬や治療を行おうとしている.

本研究では, 遺伝子多型であるSNPと生活習慣病の関連性を調査し, 生活習慣病の発症危険性を内科医に提示するシステムの構築を目的とする. このシステムを用いることで, 内科医は患者に対し, より的確に生活指導を行うことが可能となる.

本論文は以下の様に構成される. 第2章では, 研究の対象となる遺伝子とその遺伝子が持ち得る情報について説明し, 第3章で既存のシステムの現状の問題点を述べる. 第4章で提供されたデータを分析し, 第5章と第6章でシステムの構成を説明する. 第7章に行った実験とその考察を述べ, 第8章でまとめる.



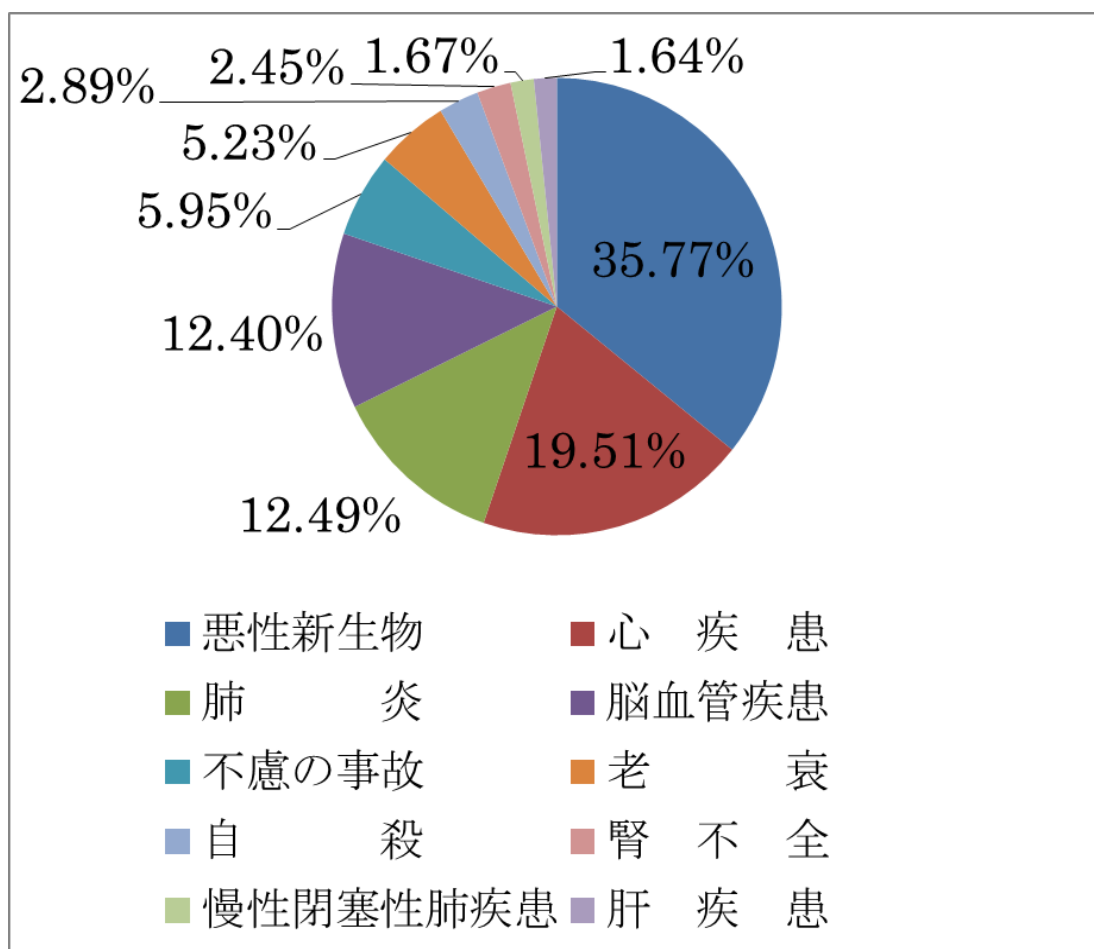


図 1-1 日本人の死因割合（10万人対）（2011）

## 第2章

### 遺伝子情報のSNP

.

生物はすべての細胞内に染色体をもっており，その染色体を構成するものが遺伝子である．遺伝子はその生物一個体の性質の情報を有すものであり，生体の設計図の役割を果たしている[3]．

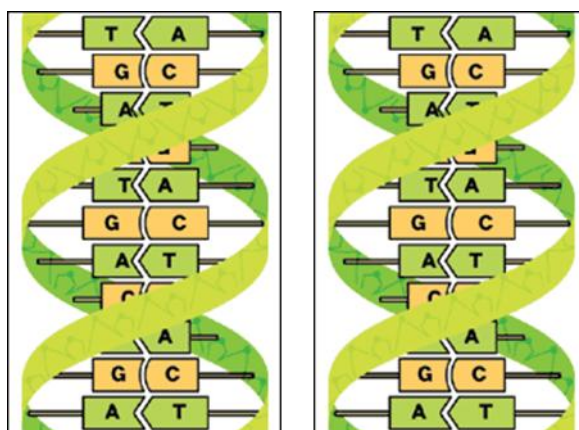


図 2-1 DNA の構造 (塩基対)

#### 2.1 塩基配列

遺伝子の性質は4種類の塩基の並び方，すなわち塩基配列によって決定される．4種類の塩基とはアデニン，グアニン，シトシンそしてチミンであり，一般的にそれらの塩基は図 2-1 のように，頭文字である A，G，C，T を用いて表現される．



図 2-1 塩基配列の一般的な表現

## 2.2 SNP (スニップ)

ヒトの塩基配列はそのほとんどが同じである。しかし、個体によって差異が現れる特定の箇所が存在する。その箇所のことをSNP(スニップ)と呼んでいる。SNPとは一塩基多型(Single Nucleotide Polymorphism)の略称である。

ヒトの場合、SNPは300万~1000万個ほど存在するといわれており、その数は遺伝子全体の約0.1%に当たる。

SNPがその個体自身の体質を決定することは確実視されている。よって、病気のかかりやすさ、あるいは薬の効きやすさ等はこのSNPにより決定される。そのため、SNPを解析することにより、個人の体質を知ることが可能となる。

また、ヒトは同じ遺伝子の塩基配列を2つ持っており、SNPは一か所につき2つ存在することになる。そこで、SNPはA/AやC/T,G/Gのように二つの塩基の組み合わせによって表現される。

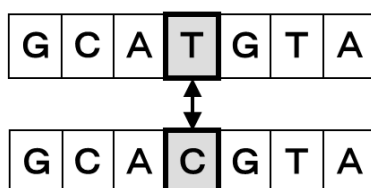


図 2-2 SNP

## 2.3 SNPの違いによる体質の変化

前節にて、SNPにより個人の体質が決定することを述べた。この節では、その具体例を一つ紹介する。

通称、酒酔い遺伝子と呼ばれるSNPが存在する。このSNPによりヒトのアルコールに対する強さが決定される。図2-3に示すようにこのSNPがG（グアニン）の場合酒に強く、A（アデニン）の場合酒に弱くなる。

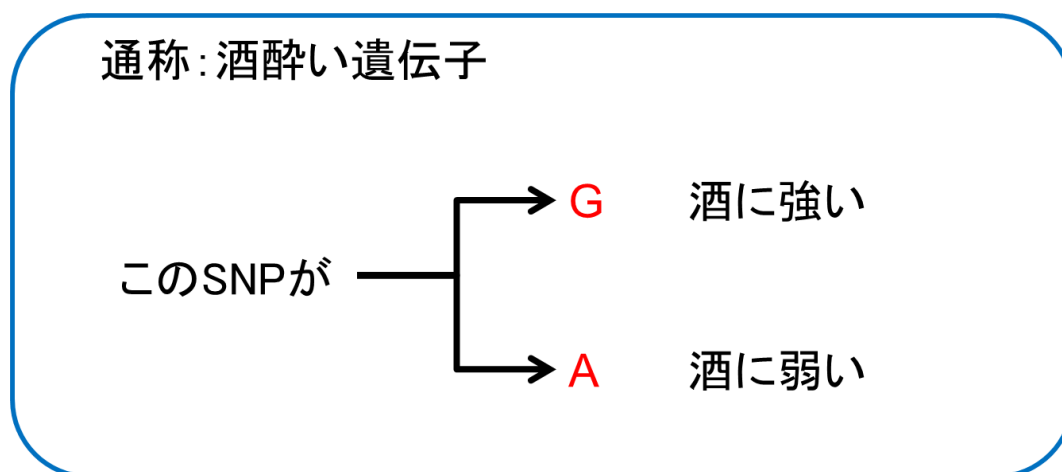


図2-3 SNPによる体質の違い（一例）

## 第3章

# 既存の健康診断支援システム

### 3.1 健康診断支援システムの必要性

生活習慣病とは、生活習慣がその発症・進行に関与する疾患群のことである。生活習慣病を放っておくと重大な病気や様々な合併症を発症する。また、重度の生活習慣病になると、完治することは難しく、厳しい食事・運動制限が課せられる。そのため、かかる前あるいは軽度のうちからの予防・治療が重要となる。

生活習慣病は、一つのSNPに直結して発症するのではなく、いくつかのSNPと血糖値、血圧、コレステロール値等の生体指標が絡み合っ、発症すると考えられている。そのため、SNPsと生体指標の状況から、発症のリスクを予測することが可能であると考えられている。

本研究で開発するシステムは遺伝子と生体指標症の両方を考慮し、発症の危険性を推定するものである。

### 3.2 既存の健康診断支援システム

SNPsの情報を利用した健康診断支援システムとして、三重大学生命科学支援センターの山田芳司教授のアイデアをシステム化した、G&Cサイエンス社より提供されている『genomarker』というシステムが存在する[4][5]。このシステムは病院にて採取した特定のSNP情報をもとに、被験者の生活習慣病発症リスクを提示するものである。

しかし、分析に用いている手法が古いものである、保険が適用されない、出力が利用しにくい等の問題から導入している病院は少ない。

そこで本研究ではより高いレベルの分析による、利用しやすい形態のSNPsを用いた健康診断の支援を行うシステムの開発を目指している。

### 3.3 Genomarker での検査方法

『genomarker』における検査の流れを以下に示す。

①検査可能な医療機関にて検査の説明を受ける

②説明を受けたのち同意書等を記入

③検体の採取

血液または口腔粘液を 1ml 採取

検査結果は約 2 週間後に報告される。検査結果について担当医師からの説明を聞くことで生活習慣改善の足掛かりとすることができる。

### 3.4 Genomarker の検査結果

検査可能である項目を表 3-1 に示す。また、検査結果の一例を図 3-1 に示す。検査結果には発症リスクを数値化したものに加え、日本人の平均に対してどの程度かかりやすいのかを示すオッズ比が記載されている。また、簡単な生活指導も記載されている。

表 3-1 Genomarker で検査可能な項目

肥満
メタボリックシンドローム
糖尿病
高血圧
慢性腎臓病
心筋梗塞
脳梗塞
脳出血
くも膜下出血
ミトコンドリアハプログループ

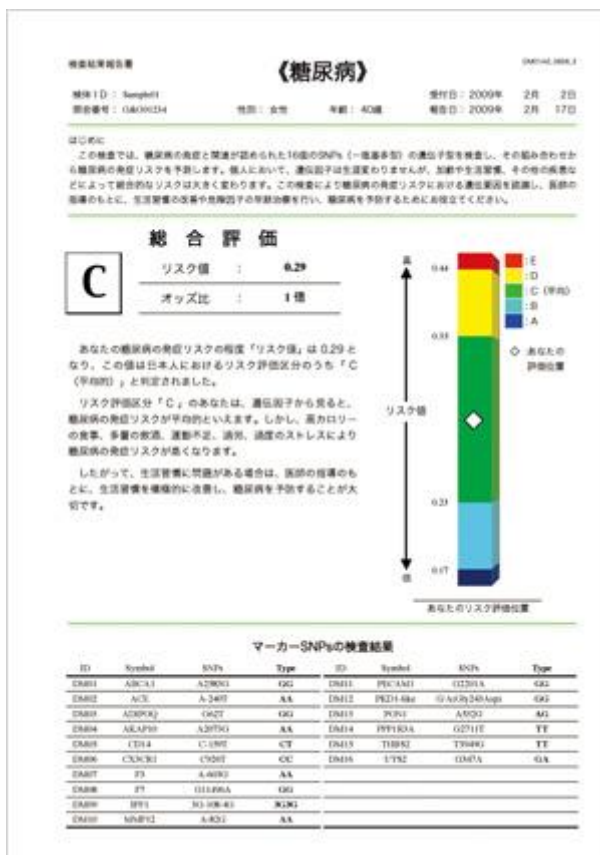


図 3-1 糖尿病の診断結果（例）



## 第4章

# データの分析

精度の高い発症危険度算出手法の検討を行うためには、SNPsと生体指標がどのような特徴を有しているか調査する必要がある。この章では、システム作成に当たり、事前に行った簡単な分析の結果を示す。

### 4.1 遺伝子・生体指標集積データ

提供されたデータには、生活習慣病に関係すると予想される70か所分のSNPの塩基に加えて、各種生体指標、生活習慣病の発症、非発症が6853名分記載されている。表4-1にデータに記載されている生体指標と生活習慣病の種類を示す。

また留意する点として、これらのデータは日本の病院に来院した方々の協力により完成したものである。したがって、データ採取協力者の年齢に大きな偏りがある、日本人があまり持っていないSNPタイプのデータが少ないなどの一般性を欠いた部分が見受けられる。そのためシステムの構築・実験の際には、この点に十分注意し検討を行う必要がある。

表 4-1 遺伝子・生体指標集積データに記載されている情報

生体指標	生活習慣病
肥満指数 (BMI)	高血圧
年齢	高脂血症
性別	糖尿病
喫煙の有無	慢性腎臓病
最高血圧	急性心筋梗塞
最低血圧	
総コレステロール値	
HDL-C	
LDL-C	
中性脂肪	
血糖値	
糖化ヘモグロビン量	
血清クレアチニン濃度	
腎臓の糸球体濾過量	

※ BMI= 体重[kg] / 身長[m] × 身長[m]

HDL-C:善玉コレステロール

LDL-C:悪玉コレステロール

## 4.2 データの分布

発症危険度の算出を行うためには、発症者と非発症者の間に特徴量の違いがなくてはならない。そこで、任意の2特徴量を軸として発症者と非発症者および、あるSNPの種類分布状況を調査した。図4-1, 図4-2にそれらの分布図を示す。

図4-1, 図4-2からは、ともに分布の違いを確認できる。しかし、分布形状の違いはあるものの、重複している箇所が多いことが判明した。したがって、これらのデータから発症危険度の算出を行うためには、分布の代表点を利用するだけでなく、分布形状も結果に反映できる方法を取る必要性がある。

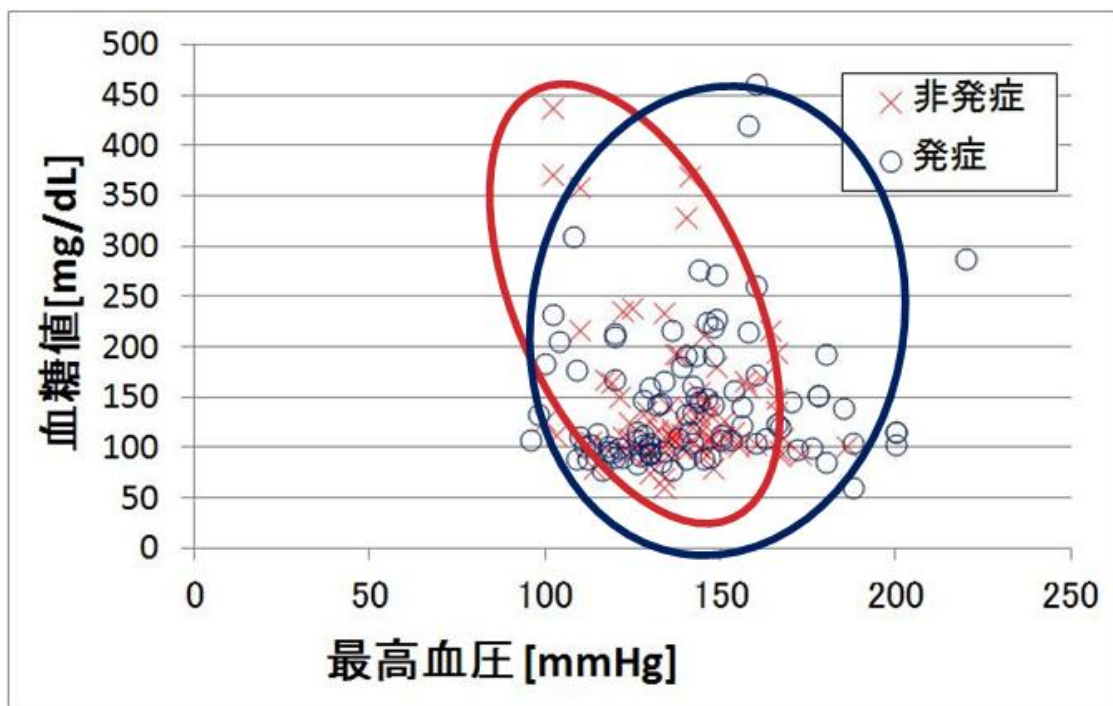


図 4-1 発症者の分布状況

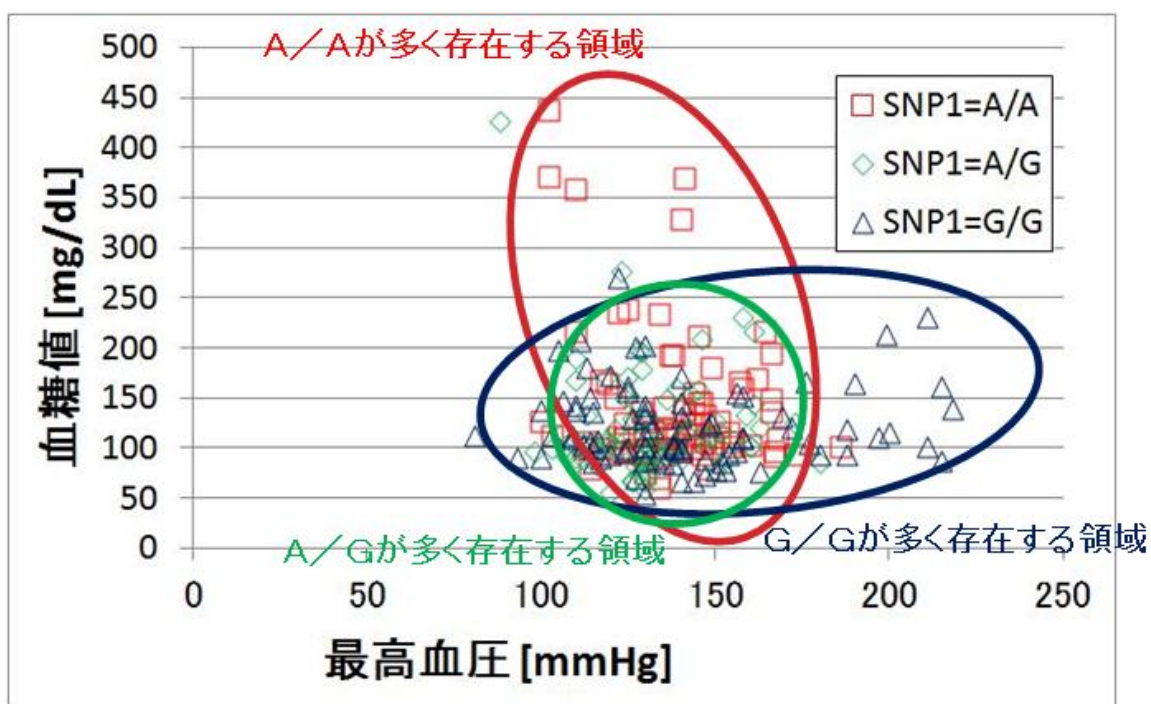


図 4-2 ある SNP の分布状況

## 第5章

### 提案システム

#### 5.1 提案システム

本研究において構築するシステムは、図 5-1 に示すように、医師が患者の遺伝子と生体指標データを入力することにより、患者独自の生活習慣病発症の危険性を推定し出力するものである。このシステムにより内科医は、患者に対してよりの確に生活習慣改善の指導を行うことができる。このシステムの構築には Windows 上の R 言語を用いた[6]。R 言語は遺伝子解析用のライブラリを有しており、また統計的処理を行う関数も多い。そのため、解析と出力を同時に行う提案システムの構築に適していると判断した。

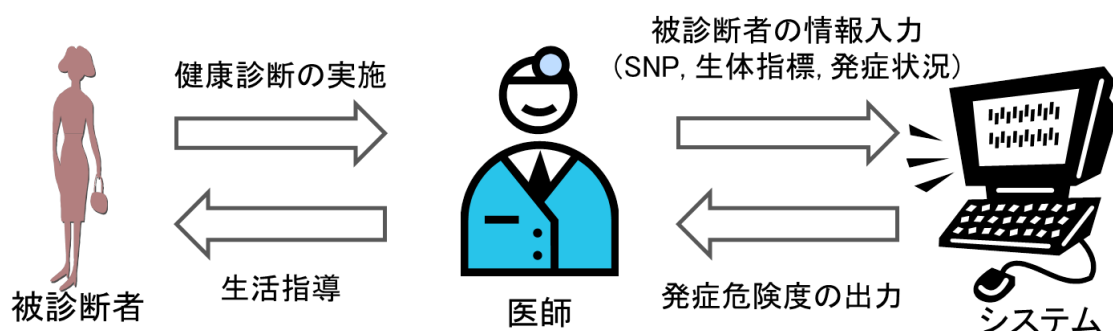


図 5-1 提案システムの構想

## 5.2 的確な生活指導を実現する出力

既存のシステムでは、被験者に対して遺伝的な発症危険度のみを提示していた。しかし、生活習慣病発症の危険性は生体指標値の増加に伴い、遺伝子ごとに異なった増加傾向を示すと考えられる。したがって、遺伝子のみにより算出された発症危険度では、的確な生活指導を行うことができない。そこで、提案システムの理想的な出力は図5-2に示すような、その患者独自の発症危険度を推定した発症危険度推定曲線とする。この曲線を用いることにより、現在の発症危険度を知るだけでなく、生活指導によりどのような効果が得られるのかを被験者に示すことができる。発症危険度とは、SNPや現在の健康状態を加味し、生活習慣病を発症する危険性を数値化したものである。

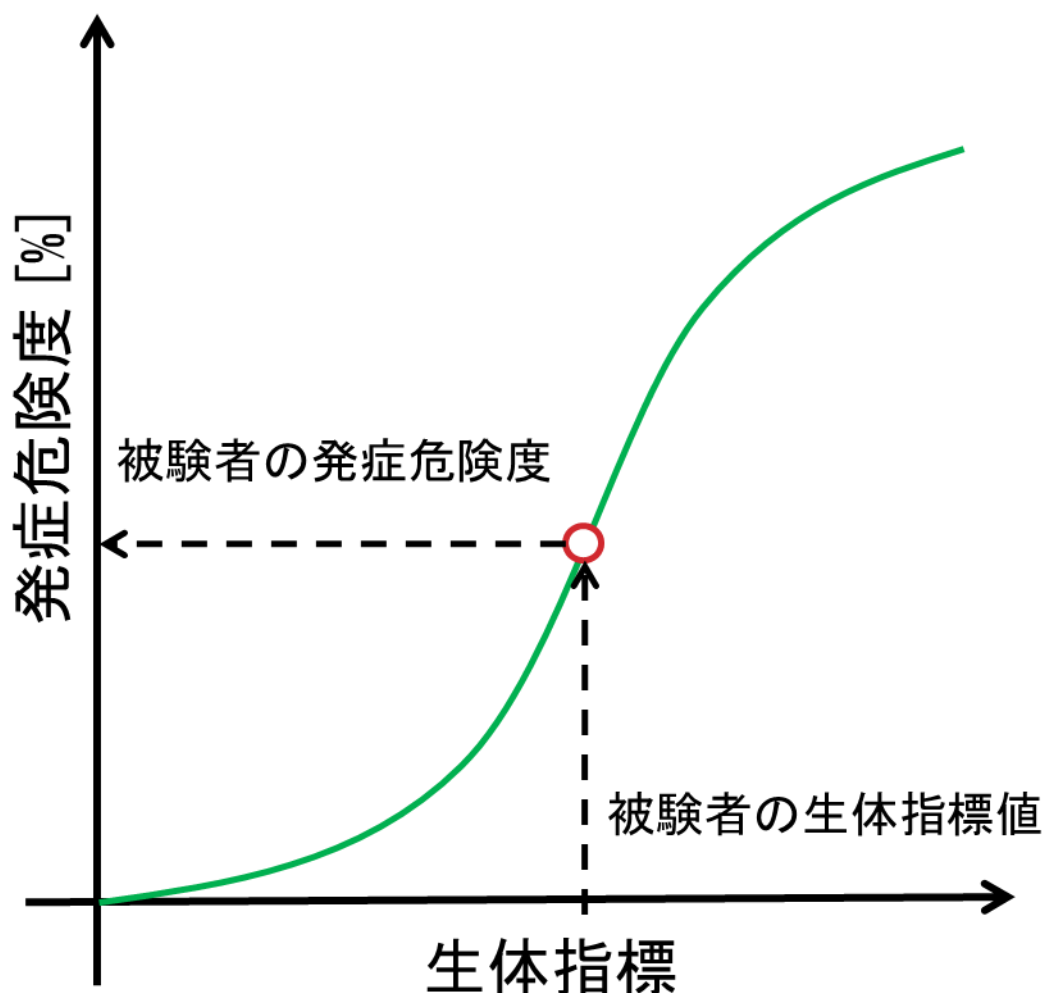


図5-2 提案システムの出力（発症危険度推定曲線）

## 第6章

### 提案システムの概要

提案システムは大きく二つの部分により構成される。一つは、医師から事前に提供されているデータ群より統計的に生活習慣病発症の特徴を分析し、診断に用いる辞書を作成する過程。もう一つは、被験者のデータを入力した際、前述の過程において作成した辞書と照らし合わせ、発症の危険性を計算する過程である。

本研究の流れを図6-1に示す。この章では6.1節で診断用辞書を作成する学習過程を、6.2節で発症の危険性を算出し出力情報を得る健康診断過程を説明する。

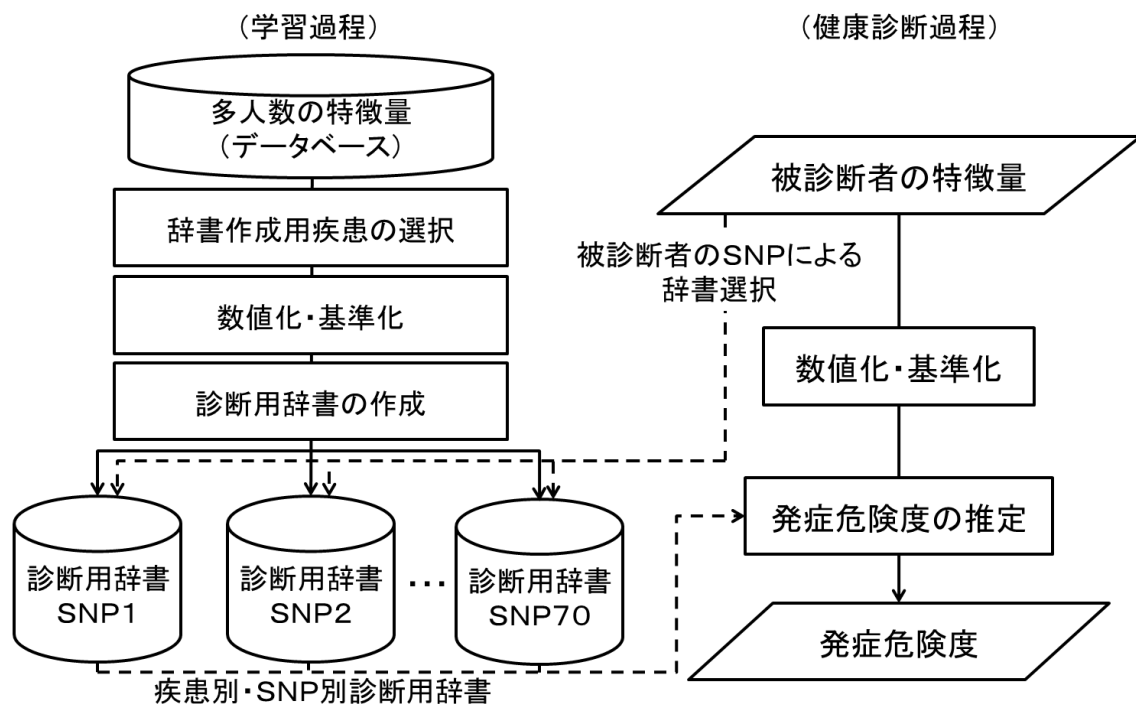


図6-1 研究の流れ

## 6.1 学習過程

提供された遺伝子と生体指標のデータをもとに，SNPおよび生体指標と生活習慣病の関連性を示す指標を作成し，システム内で辞書として扱う．

### 6.1.1 辞書作成用疾患の選択

提案システムでは，発症危険度を疾患ごとに算出する．そのために，診断用辞書も疾患ごとに用意しなければならない．そこで，診断用辞書の作成を行う際にはどの疾患に対して辞書を作成するか決定する必要がある．

3.1.1 節においてあげた疾患から一つを選択する。次節以降の過程ではこの過程で選択された疾患についての辞書作成を行う．



### 6.1.2 データの基準化

辞書を作成する前に、分析を正しく行うためにデータに前処理を行う必要がある。

SNPデータの各箇所は2種類の塩基にそれらが組み合わさったものを加えた3種類のいずれかとなる。そのため、SNPには表6-2に示すように一箇所につき、それぞれ1,2,3の値を与える。

また、特徴ベクトルの各要素は、スケールがそれぞれ異なっている。そこでスケールを統一させるため(1)式により基準化を行う[7]。この操作により分布形状が整えられ、計算可能な形となる(表6-2)。基準化されたデータは、各値が要素となる特徴ベクトルとなり、一人分のデータから一つの特徴ベクトルが作成される。

表6-1 SNPの変換ルール(例)

SNPの種類	変換後の値
A/A	1
A/G	2
G/G	3

$$\hat{x} = \frac{x - \bar{x}}{\sigma} \quad (1)$$

$\hat{x}$ : 基準化された特徴量

$x$ : 特徴量の値

$\bar{x}$ : 平均値

$\sigma$ : 標準偏差

表 6-2 基準化後の値 (例)

要素	計測値	基準化後の値
SNP1	1 (A/A)	0.281
SNP2	3 (T/T)	-0.25
SNP3	2 (A/G)	0.002
・	・	・
・	・	・
血糖値	97	0.81
最高血圧	130.9	1.22
・	・	・
高血圧	0	0.23
・	・	・

### 6.1.3 診断用辞書の作成

個人が有する可能性のあるSNPの組み合わせは非常に多い。そのため、すべての組み合わせに対して辞書を作成することは困難である。そこで、SNPごとに種類別の辞書を作成し、被験者のSNP情報に対応した辞書を選択することにより、疑似的に被験者の持つSNPに応じた辞書を作成する。

そのために辞書作成用のデータを各SNPの種類ごとに分類する。各SNPは1か所につき3種類あるので、SNPの数(70か所)×3種類の210パターンにデータは分類される。

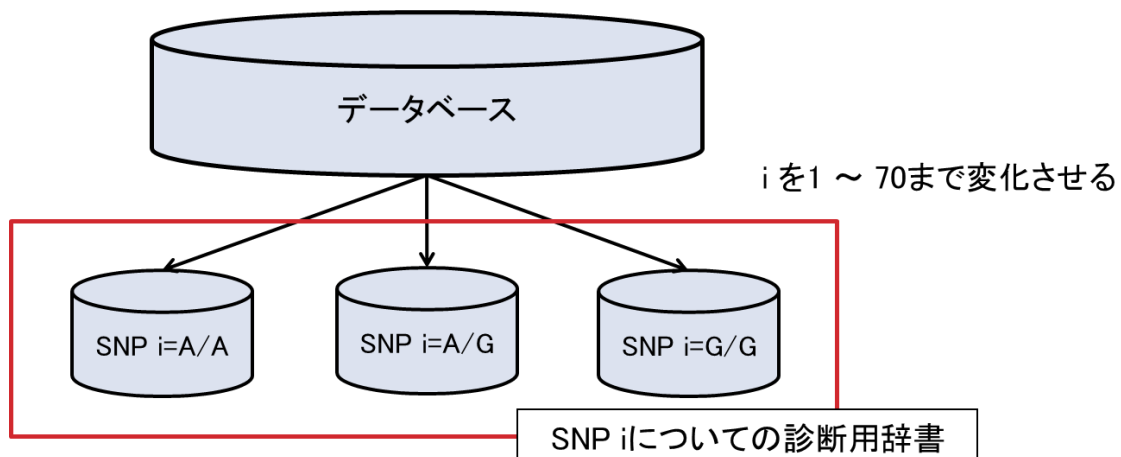


図 6-2 診断用辞書作成用データの分類

次に特定疾患の発症者非発症者の群を形成するため、図6-3のように、患者が特定疾患を発症しているかどうかで患者を分類する。この分類により多次元特徴空間内には発症者、非発症者のベクトル群が形成される。各々の群で平均化を行うことで、特定疾患発症者と非発症者の特徴平均ベクトルを作成する。特徴平均ベクトルは6.1.3節において分類されたデータセットごとに作成される。これらの特徴平均ベクトルとその共分散行列は発症、非発症の特徴であるため、診断用辞書として発症危険度の計算に用いる。これにより、図6-5のような疾患別・SNP別の診断用辞書が作成される。

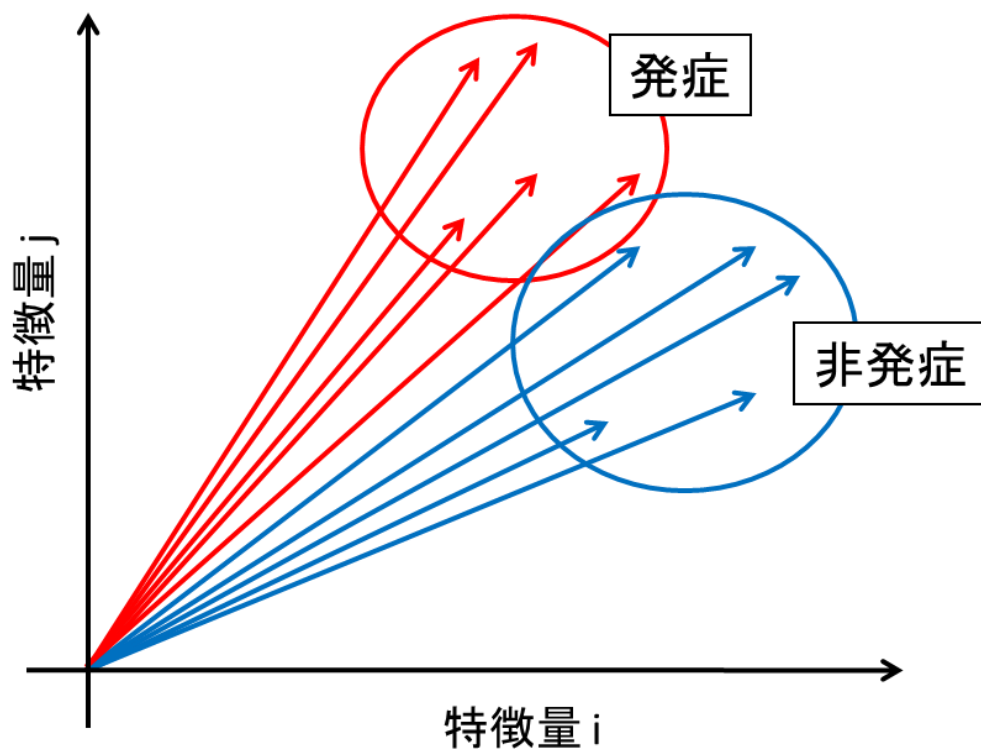


図6-3 特徴ベクトルの分類 (サンプルの分布イメージ)

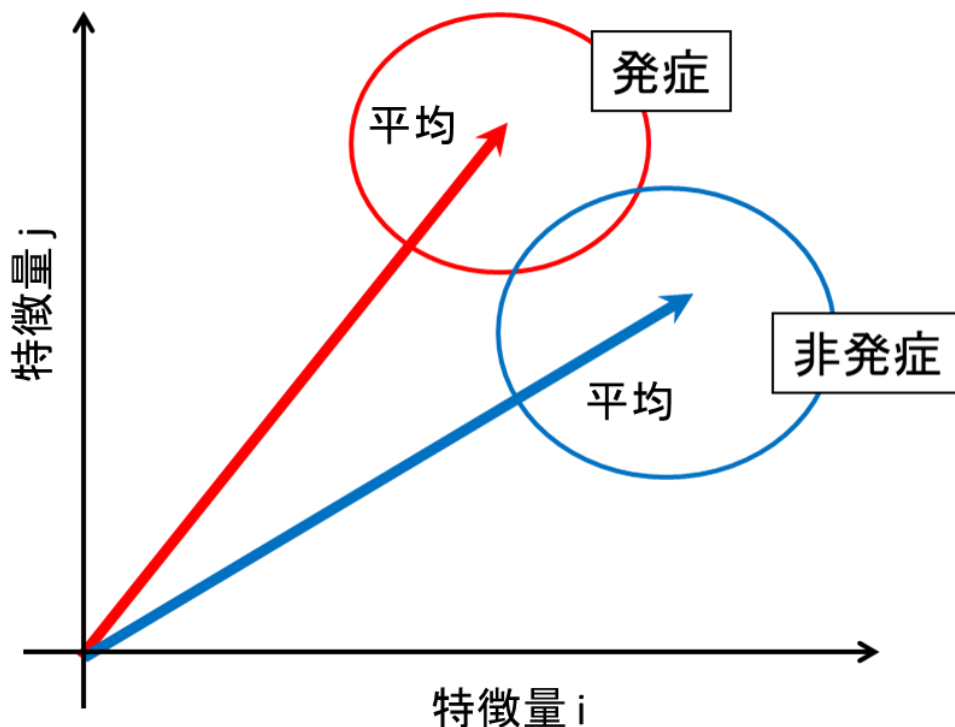
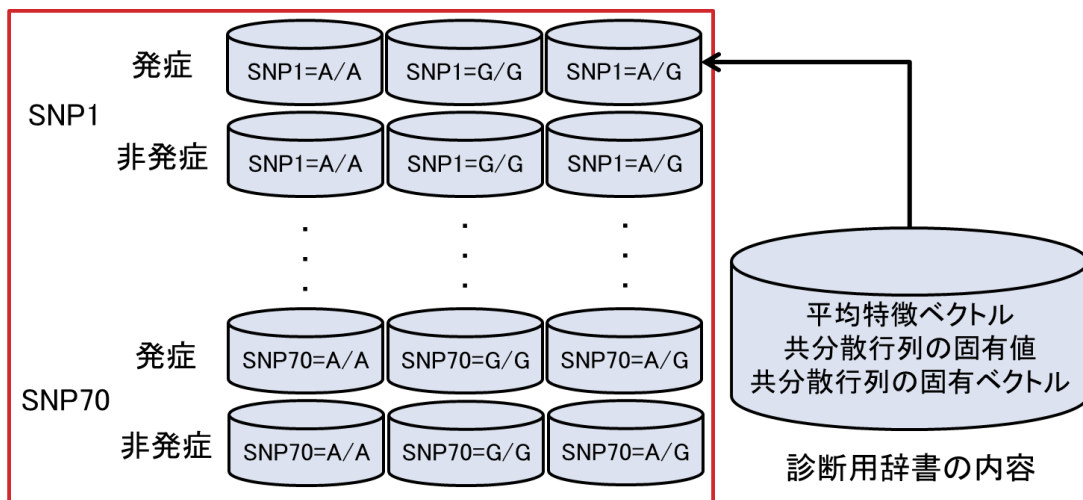


図 6-4 特徴平均ベクトルの作成 (イメージ)



疾患別・SNP別診断用辞書

図 6-5 疾患別・SNP別診断用辞書

## 6.2 健康診断過程

この節では被験者のデータを入力した際に、発症危険度の算出と発症危険性推定曲線を作成する方法について述べる。

### 6.2.1 診断用辞書の選択

被験者のSNP情報をもとに、診断用辞書から個人のSNPに応じた診断用辞書の選択を行う。このように選ばれた辞書を用いることで、疑似的に被験者と同じSNPの組み合わせを持った辞書との比較を行うことが可能になる。

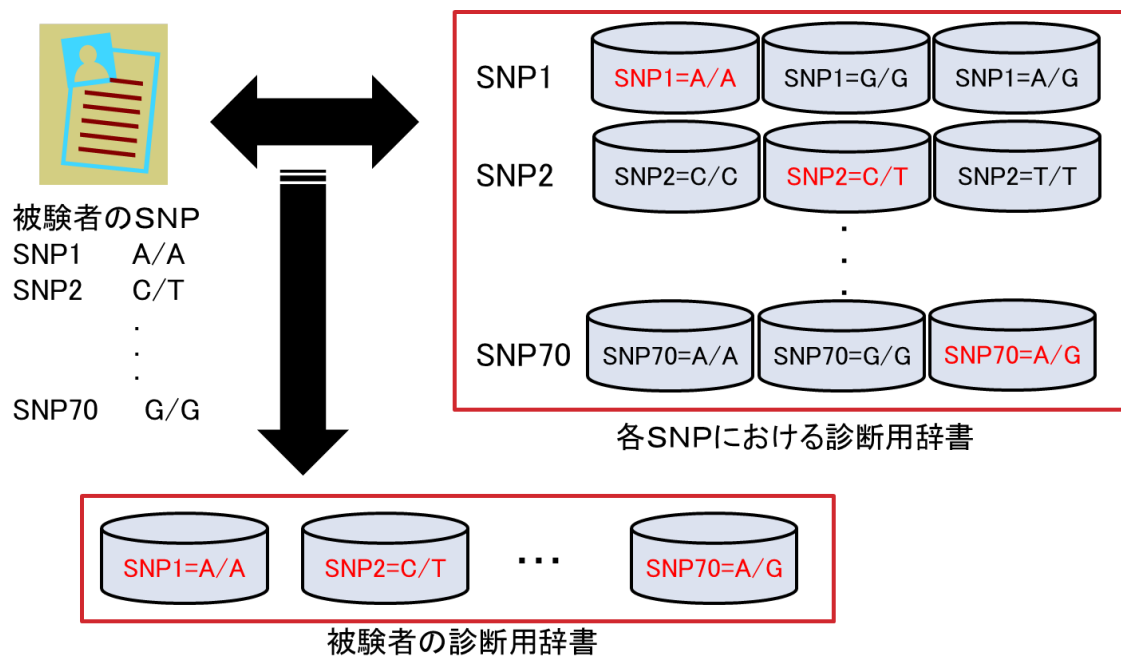


図 6-6 被験者の診断用辞書の選択

## 6.2.2 発症危険度の計算

システムは入力された、被験者一人分の遺伝子・生体指標データから 6.1.3 節と同様の方法により入力ベクトルを作成する。この入力ベクトルと 6.1 節で作成された発症者・非発症者の特徴平均ベクトルとの距離から、発症危険度を計算する。図 6-7 はその様子を示している。

発症危険度を正しく算出するには要素各々の重みと分散を考慮した精度の高い識別が必要になる。そこで、本研究では(2)式に示す疑似ベイズ識別関数を用いてベクトル間の距離を算出する[7]。

(2)式により計算された距離をもとに(3)式によって発症の危険度を計算する。

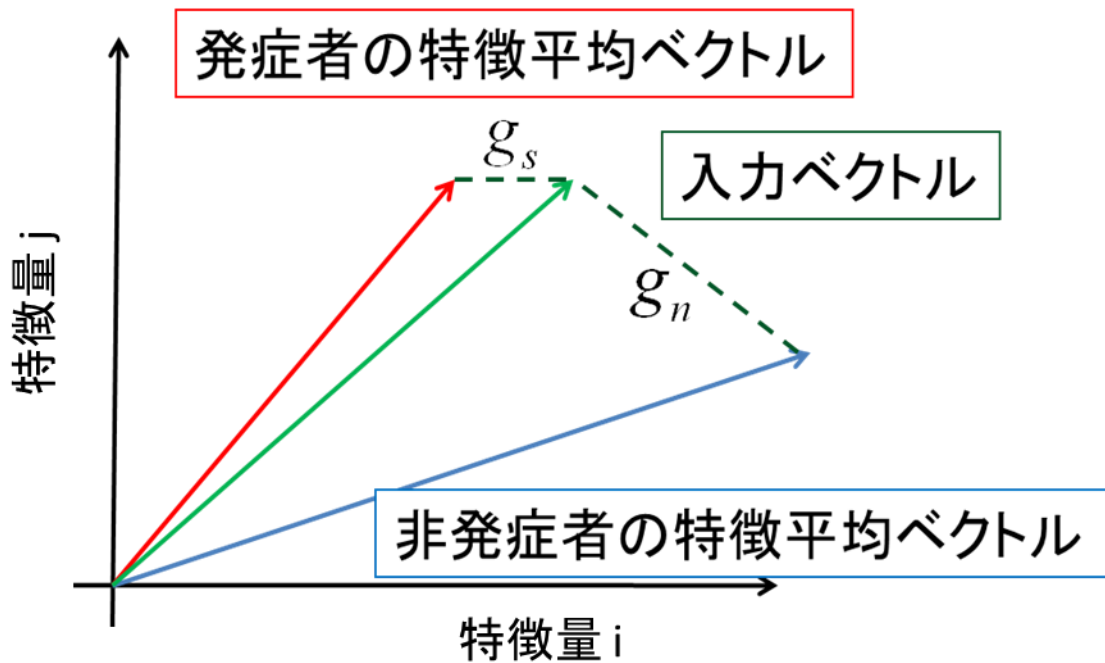


図 6-7 入力ベクトルとの比較 (イメージ)

$$g(x) = \frac{1}{\lambda_{k+1}} \left\{ \|x - \mu\|^2 - \sum_{i=1}^{k+1} \left(1 - \frac{\lambda_{k+1}}{\lambda_i}\right) (x - (\mu, \phi_i))^2 \right\} + \ln\left(\prod_{i=1}^k \lambda_i \prod_{i=k+1}^n \lambda_{k+1}\right) \quad (2)$$

$\mu$  : 特徴平均ベクトル

$\lambda$  : 特徴平均ベクトルの共分散行列の固有値

$\Phi$  : 特徴平均ベクトルの共分散行列の固有ベクトル

$(\mu, \Phi)$  :  $\mu$  と  $\Phi$  の内積

※第2項  $+\ln(\prod \lambda \prod \lambda)$  はサンプル数が少ない場合には

固有値が正しく推定できないので同じ値を使用することを意味する。

※今回のシステムでは  $k=20$  とした。

$$\text{発症危険度} = \frac{1}{70} \sum_{i=1}^{70} \frac{\exp(g_{ni})}{\exp(g_{ni}) + \exp(g_{si})} \times 100 [\%] \quad (3)$$

$g_n$  : 非発症者の特徴平均ベクトルとの距離

$g_s$  : 発症者の特徴平均ベクトルとの距離

$i$  : SNP の番号



### 6.2.3 発症危険性推定曲線の作成

6.2.2 節により算出した発症危険度より、システムの出力となる発症危険性推定曲線を作成する。

まず発症危険度推定曲線の横軸を選択する。その際、横軸は特定疾患の発症に大きくかわる要素とする。つまり、出力の発症危険度推定曲線はここで選ばれた要素が変化したときに発症危険度がどのように変化したかをシミュレートしたグラフとなる。

発症危険度は一つの入力に対して一つの値が計算される。そこで、横軸として選択された要素を細かく変化させ、その都度 6.2.1 節の計算により発症危険度を算出する。すると、横軸の変化に対する発症危険度の値が細かく計算されることになる。それらの発症危険度を縦軸が発症危険度、横軸が選択された要素のグラフ内ではつなげると発症危険度の推定曲線を得ることができる。その際に、選択された要素の変化範囲は一般的な低値（正常域）から高値（異常域）とし、変化させる幅はできる限り細かいものとする。

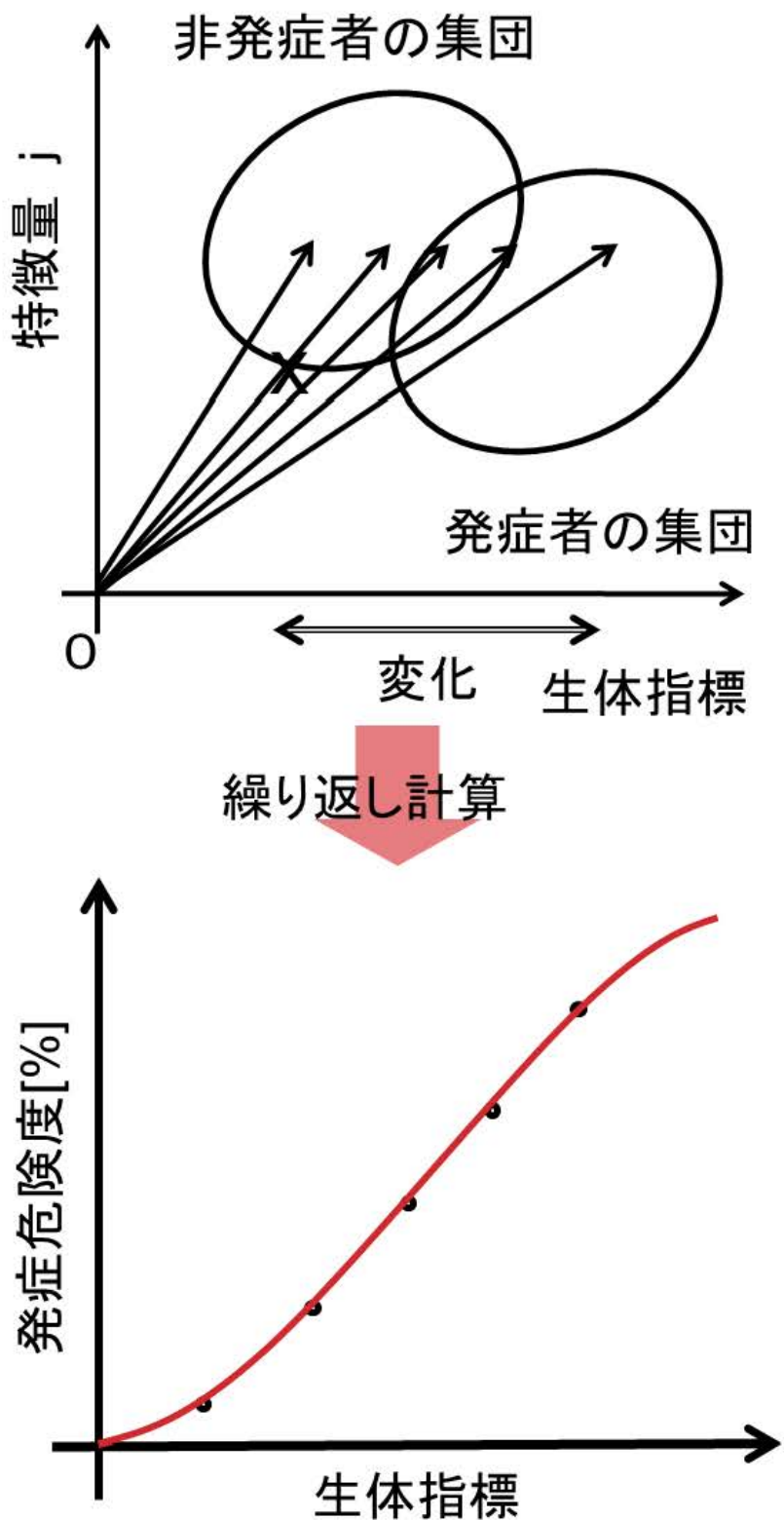


図 6-8 発症危険度推定曲線の作成

## 第7章

### 評価実験

提案システムにおいてSNPに対する発症危険度の性能を検証する。そこでSNPのパターンが異なる被験者ごとに発症危険度の計算を行い、発症危険性推定曲線を作成し、発症危険度がどのように変化するかを調査した。

#### 7.1 実験1 -SNPによる発症危険度の違い-

##### 7.1.1 実験方法

生活習慣病の発症には、生体指標が大きくかかわっていることはよく知られている。そのため、単純に提案システムの出力結果を示してもSNPによる影響を反映できているかどうかを検討することができない。そこで実験1では、SNPのみによる発症危険度の違いを観測するため、評価用の被験者はSNP以外の要素を統一した。実験に用いた被験者のデータはSNPの組み合わせが異なる6名分である。統一した値を表7-1に示す。

本システムの出力を確認するために、特定疾患を急性心筋梗塞として評価実験を行った。診断用辞書の作成には発症者1597人、非発症者713人を用いる。

発症危険度推定曲線を作成する際の横軸として、急性心筋梗塞を引き起こす一要因として考えられている最高血圧値を用いる。最高血圧値は100～200[mmHg]の範囲で変化させる。

表 7-1 統一した値

要素	計測値	基準化後の値
BMI	25.9	0.44
年齢	43	-2.7
性別	1(男)	0.82
喫煙状況	1(喫煙)	1.6
最低血圧	68	-0.85
総コレステロール	245	1.7
HDL-C	189	-0.75
LDL-C	42	2.1
中性脂肪	245	0.45
血糖値	91	-0.73
HbA1c	5.4	-0.31
血清クレアチニン	0.72	-2.0
糸球体濾過量	915	1.2
高血圧	0	-1.1
高脂血症	1	1.4
糖尿病	0	-0.67
慢性腎臓病	1	-0.66

## 7.1.2 実験結果

図7-1 に出力された結果を示す。

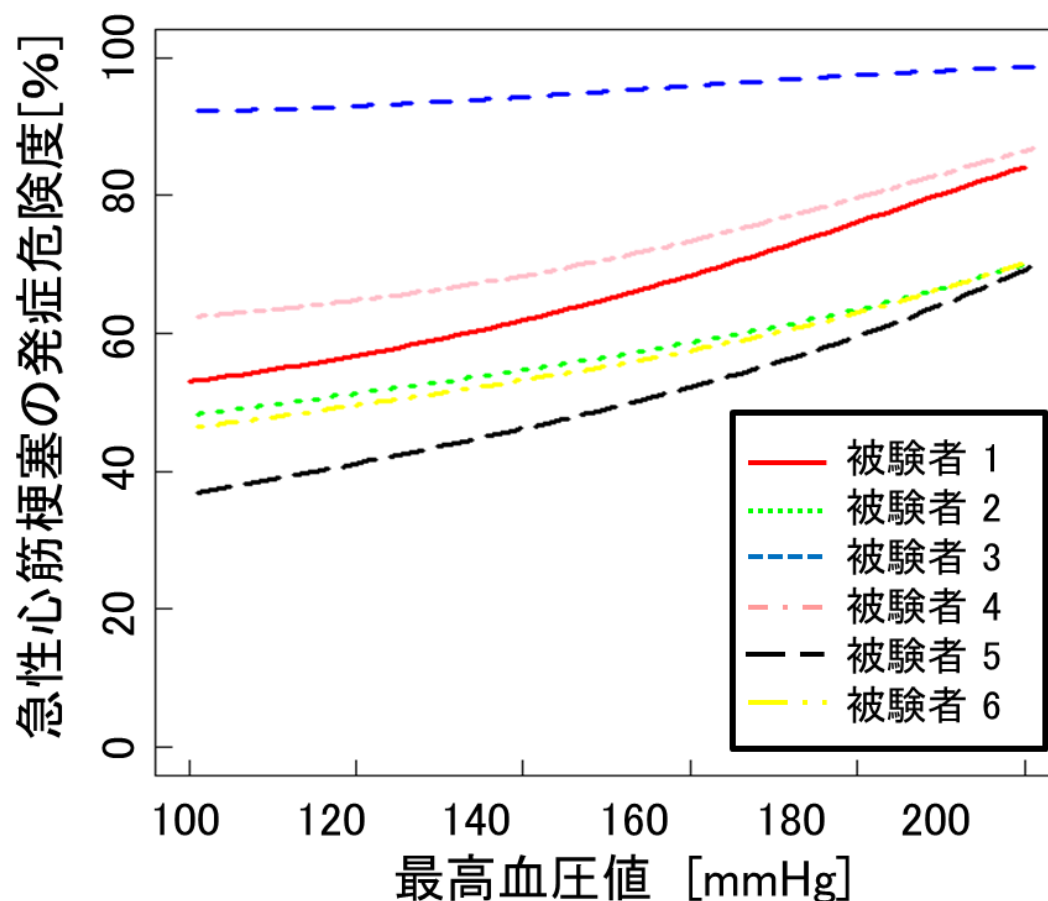


図7-1 実験結果

実験の結果より、被験者ごとに発症危険度の値・増加傾向が異なることが確認できた。全被験者でSNP以外の値が統一されていることから、これらの違いはSNPによるものであると断定できる。

また、最高血圧値の変化に伴い発症危険度の値が増加していることから、急性心筋梗塞の発症に最高血圧値が大きくかかわっていることも確かめられた。

## 7.2 実験2 —個人の発症危険度推定曲線—

実験1ではSNPによる発症危険度の違いのみを確認した。実験2では実際に被験者の発症危険度を算出する場合とおなじ条件で発症危険度の計算を行う。

### 7.2.1 実験方法

特定疾患，診断用辞書，発症危険度推定曲線の横軸は実験1と同一のものを用いる。また，本実験では多くのサンプルを用いて，性能を検証するため，特定疾患の発症者50名，非発症者50名の計100名において実験を行った。

## 7.2.2 実験結果

図 7-2, 図 7-3 に出力された結果を示す. 図 7-2 は特定疾患の発症が確認されていない被験者, 図 7-3 は特定疾患を発症していることが確認されている被験者について作成された発症危険度推定曲線である.

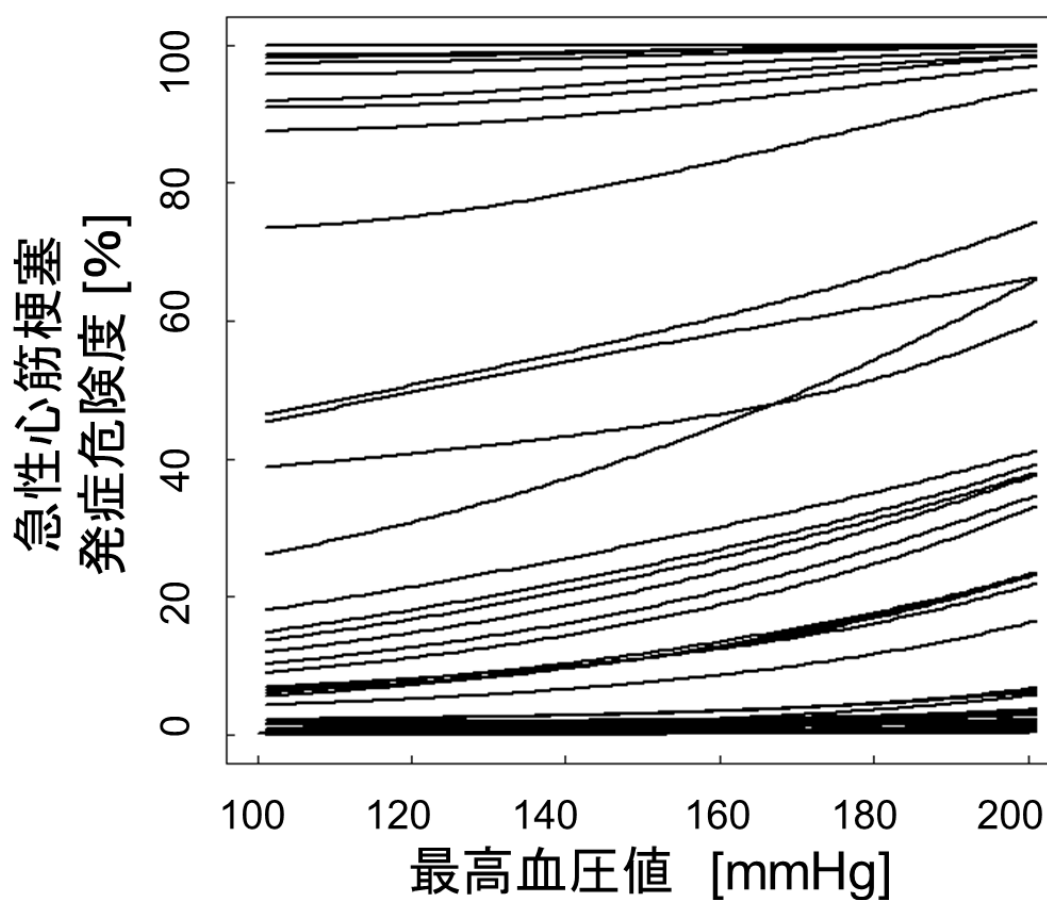


図 7-2 非発症者 50 名の発症危険度推定曲線

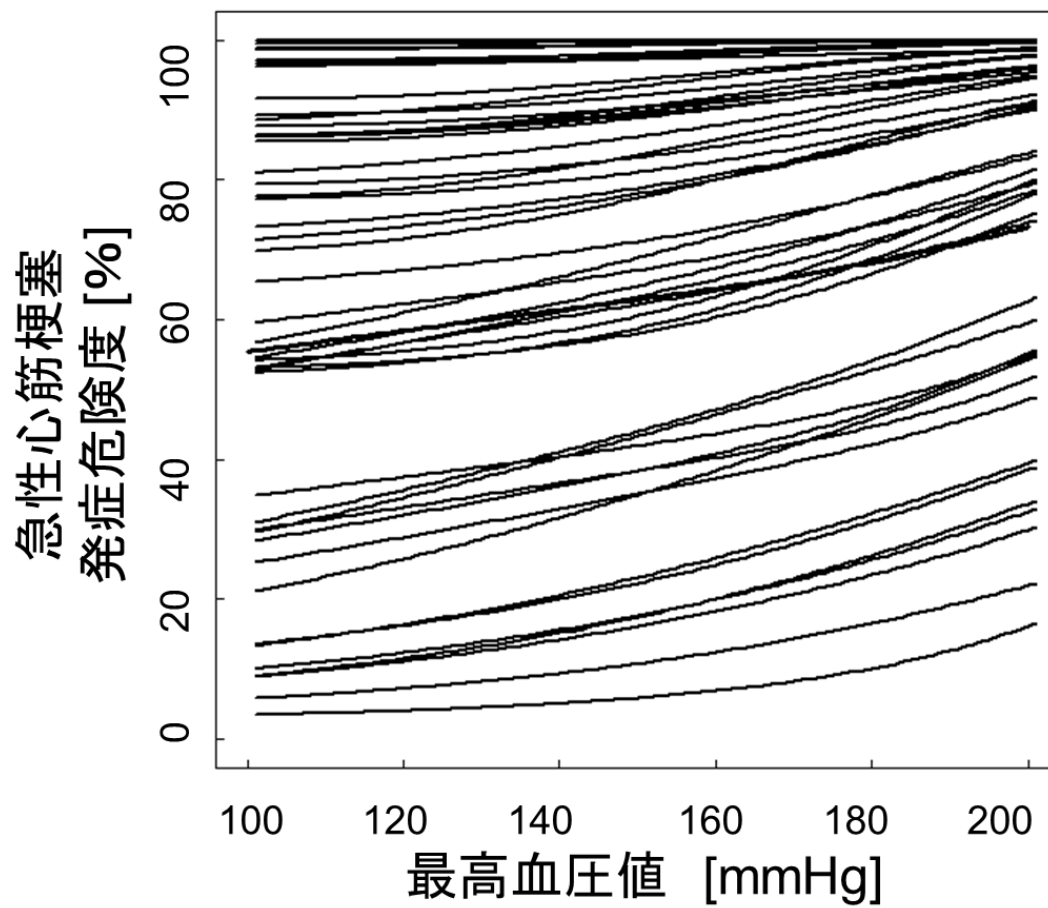


図 7-3 発症者 50 名の発症危険度推定曲線



### 7.3 実験の考察

実験1の結果より、提案したシステムは SNP における発症危険度の違いをとらえることができるものであることが判明した。しかし実験2では発症危険度の変化が小さい被験者が存在することも判明した。特に、非発症者においてその傾向は顕著である。これは、実験に用いたデータが採取した環境により非健常者に大きく偏ってしまっていたことが原因であると考えられる。

この出力における問題として、実験1の被験者3のように、発症危険度の値の変化が小さく、かつ極めて高い（あるいは低い）値になってしまっている被験者が多いことがあげられる。SNPの値により個人差があるとはいえ、生体指標が正常な状態で高い発症危険度を示すことは一般的に考えられず、発症危険度の推定が正しく行われていないと考えられる。

提案システムにおいて利用している方法は、発症者と非発症者の分布の違いを計り、発症危険度を数値化するものであり、手法的に発症危険度の推定ができないとは考えにくい。そのため、このような結果が出力された大きな原因としては3.1節で述べた、データの特異性が関与していると考えられる。すなわち、来院された被験者から採取されたデータを集積したために健常な被験者のデータが少なく、偏った辞書が作成されてしまった可能性が高い。

また、他の原因として現状のシステムでは各々のSNPがもつ影響の強さを考慮していないことが考えられる。

現時点では、発症危険度の計算を行う際にすべてのSNPにおける影響を同じ重みで利用している。しかし、一般的に各々のSNPが持つ影響の大きさがそれぞれ異なることが分かっている。したがって、関係性が少ないSNPの影響が必要以上に発症危険度に反映されている可能性がある。

さらに、発症危険度の推移をシミュレートする際の問題として、最高血圧値のみを変化させたために不自然な変化になってしまったことも挙げられる。

## 第8章

### 終わりに

本稿では、SNPが生活習慣病に与える影響を考慮した生活習慣指導を目的とした新しい健康診断支援システム開発のため、SNPと生体指標を用いた多次元空間により発症の危険性を算出する方法について考案、検討している。

遺伝子解析技術の発展に伴い、遺伝子を用いた医療への期待は高まっている。従来にも遺伝子から生活習慣病の発症危険度を算出するシステムは存在していたが、問題点が多く生活指導等に使えるものではなかった。そこで、統計量の多変量解析を発達させたパターン認識の技術を利用して、集積されたデータから生活習慣病の発症にかかわる特徴を辞書化し、発症危険度の計算を行う手法について提案した。その結果、SNPと生体指標を利用した発症危険度の数値化を行うことが可能となった。また、特定の生体指標を変化させながら発症危険度を繰り返し計算することにより、発症危険度の変化をシミュレートすることも可能とし、生活指導等に利用しやすい形態となった。

今後の課題は、発症危険度の精度向上である。発症危険度の数値化は可能になったものの、計算された値に対する信用性は十分でないと考えられる。そのため、使用する特徴量を絞る、SNPの関連性に応じた重みをつける、などを行う必要がある。

## 謝辞

本研究の遂行および本稿の作成にあたり，丁寧かつ的確なご指導，ご助言を賜った三重大学地域イノベーション学研究科 鶴岡信治教授,同 坂内正明教授,同 萩原克幸教授,三重大学生命科学支援センター 山田芳司教授,三重大学工学研究科 高瀬治彦准教授,同 川中普晴助教授に心より感謝申し上げます。また，研究の遂行にあたり数多くのご助言，ご指摘を賜った情報処理研究室の皆様には感謝致します。

## 参考文献

- [1] 厚生省平成23年人口動態統計, 厚生労働省,  
<http://www.mhlw.go.jp/toukei/list/81-1.html>, 2012
- [2] 文部科学省共済組合, (監修奈良信雄), “健康診断プロジェクト”, 文部科学省共済組合, (東京法規出版), 2010
- [3] GenoMarker, G&G サイエンス株式会社, <http://www.genomarker.jp/index.html>
- [4] 永田恭介: ヒトの生物学 体の仕組みとホメオスタシス,丸善株式会社, pp.340-408 (2007)
- [5] 樋口千洋, 石井一夫: 統計解析環境Rによるバイオインフォマティクスデータ解析, 共立出版, p17-p102, (2007)
- [6] 管民郎: らくらく読める多変量解析の実践 上, 現代数学社, p167, (1993)
- [7] 鶴岡信治, 栗田昌徳, 原田智夫, 木村文隆, 三宅康二: 加重方向ヒストグラム法による手書き文字・ひらがな認識, 電子情報通信学会,論文誌 D ,Vol.7, pp.1390-1397,(1987)
- [8] James D. Watson, Tania A. Baker, Stephen P. Bell, Alexander Gann, Michael Levine, Richard Losick, 中村桂子, 滋賀陽子, 中塚公子, 宮下悦子: ワトソン 遺伝子の分子生物学, 電機大出版局, (2010)
- [9] 館野義男: バイオインフォマティクス, 裳華房, (2008)
- [10] 柳井晴夫, 高根芳雄: 現代人の統計 多変量解析法, 朝倉書店, (1979)

## 発表論文リスト

### 国内会議

- [1] 森上斐斗, 鶴岡信治, 山田芳司, 萩原克幸, 高瀬治彦, 川中普晴: 個人の遺伝子情報 (SNP) と生活習慣病の関連性を利用した健康診断支援システム, 平成 23 年度電気関係学会東海支部連合大会, A3-1, (2011)
  
- [2] 森上斐斗, 鶴岡信治, 山田芳司, 萩原克幸, 高瀬治彦, 川中普晴: 個人の遺伝子情報 (SNP) と生活習慣病の関連性を用いた健康診断支援システム, 計測自動制御学会中部支部 平成 23 年度 三重地区計測自動制御研究会講演会, (IP-06)pp1-2, (2011)

### 国際会議

- [1] Morikami, S.Tsuruoka, Y.Yamada, K.Hagiwara. H.Takase and H.Kawanaka : Health Check Support System Using Relationship between Individual Genome Information (SNP) and Lifestyle Disease, Proc. of fourth International Workshop on Regional Innovation Studies 2012 (IWRIS2012),