	12	J	5.3	5.7	5.6	5.8	67	135	244	82	5.6	528
男性計			**5.3**	**5.5**	**5.5**	**5.6**	**536**	**954**	**1806**	**745**	**5.5**	**4041**
全国平均			**5.2**	**5.5**	**5.5**	**5.6**	–	–	–	–	–	–
女性	1	K	4.9	5.2	5.4	5.5	34	64	149	50	5.4	297
	2	F	5.1	5.3	5.4	5.4	32	59	159	51	5.4	301
	3	L	5.1	5.5	5.4	5.5	73	128	284	113	5.4	598
	4	A	5.2	5.4	5.4	5.4	118	148	344	193	5.4	803
	5	I	5.1	5.3	5.6	5.5	22	37	63	45	5.5	167
	6	B	5.3	5.3	5.4	5.3	18	44	120	42	5.4	224
	7	D	5.1	5.7	5.2	5.3	19	34	88	33	5.3	174
	8	C	5.2	5.3	5.7	5.2	20	53	103	40	5.5	216
	9	G	4.8	5.2	5.4	5.6	13	56	100	45	5.4	214
	10	H	5.0	5.9	5.3	5.5	22	61	126	51	5.4	260
	11	E	5.2	5.4	5.5	5.3	27	69	125	69	5.4	290
	12	J	5.1	5.4	5.4	5.4	53	108	211	117	5.4	489
女性計			**5.1**	**5.4**	**5.4**	**5.4**	**451**	**861**	**1872**	**849**	**5.4**	**4033**
全国平均			**5.2**	**5.4**	**5.5**	**5.5**	–	–	–	–	–	–
総計			**5.2**	**5.4**	**5.4**	**5.5**	**987**	**1815**	**3678**	**1594**	**5.4**	**8074**

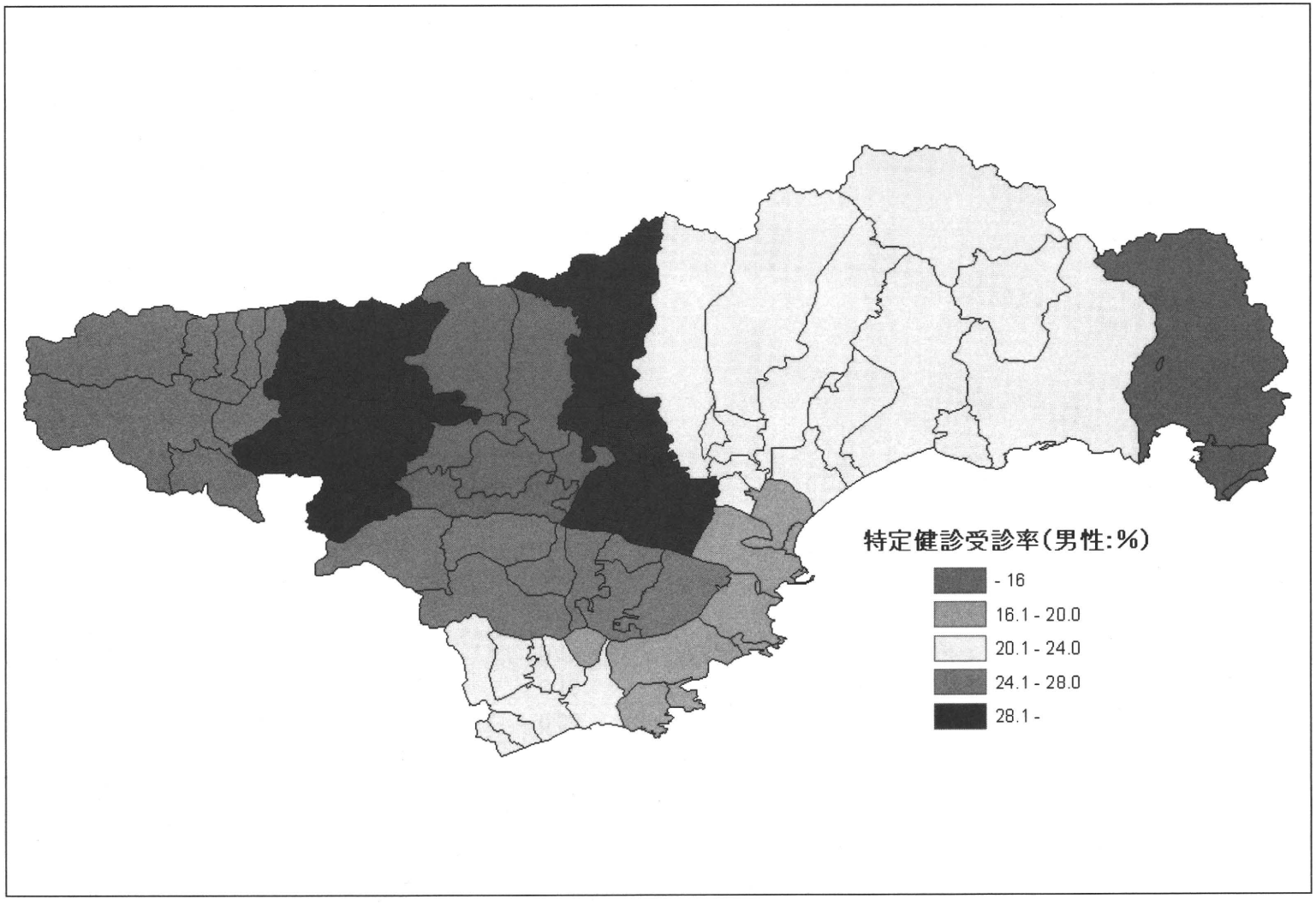
※血圧等とおなじ結果である。全国平均は「平成 18 年度国民健康・栄養の現状」より転載。

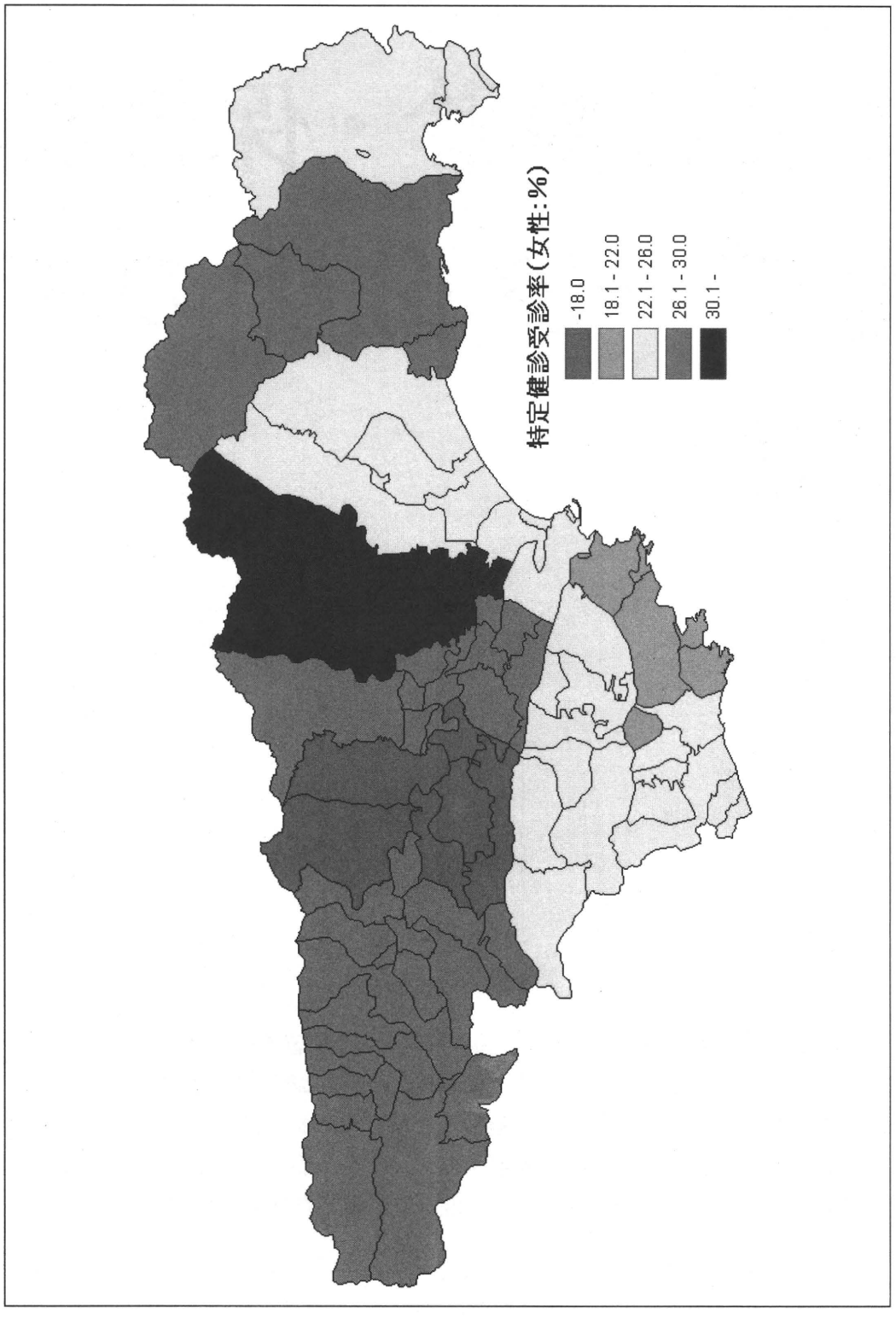
⑨健診結果の統計的分析

健康リスク	平均	標準偏差	標準誤差	95%上限	95%下限
BMI	−0.02	0.75	0.02	0.00	−0.04
腹囲	−0.29	4.45	0.11	−0.18	−0.40
収縮期血圧	−0.50	13.14	0.33	−0.17	−0.83
拡張期血圧	−0.24	7.28	0.18	−0.05	−0.42
HDLコレステロール	0.25	6.71	0.17	0.42	0.08
LDLコレステロール	−0.56	20.82	0.53	−0.03	−1.08
HbA1c	−0.04	0.52	0.01	−0.03	−0.05

※サンプル数が多く、統計的に有意差が出やすいことに留意する必要があるが、健診を連続して受けているものは健康リスクが下がっている（BMI 除く）

⑩地図




特定健診受診率(女性:%)
-18.0
18.1 - 22.0
22.1 - 26.0
26.1 - 30.0
30.1 -

【学習教材】

EXCEL・ACCESSによる健診・保健指導データ分析のてびき

平成21年度厚生労働科学研究費補助金　循環器疾患等生活習慣病対策総合研究事業
各種健診データとレセプトデータ等による保健事業の評価に関する研究班
研究協力者：藤井　仁　　国立保健医療科学院　人材育成部　主任研究官
研究分担者：横山　徹爾　国立保健医療科学院　人材育成部　部長
研究代表者：水嶋　春朔　横浜市立大学大学院医学研究科情報システム予防医学教授

1.はじめに

①対象者

公共機関にお勤めの方を想定しています。

ですので、健診データやレセプトデータはお手元にある前提で話を進めます。

②本稿の目的

本稿の目的は、日本の国保ダミーデータを用い、標準的なレセプトと健診データの処理・分析方法を学ぶことです。具体的には、できるだけ特殊なソフトを使わずに（EXCEL や ACCESS を用いて）、厚労省の「標準的な健診・保健指導プログラム(確定版)」に記載されたデータ分析をすることです。

なぜ国保データなのでしょうか？

理由は 3 つあります。

第一に、各都道府県で手に入れやすいことです。

健保や共済のデータは都道府県にまたがっており、特定の団体しか持っていないことが非常に多いです。これに対し、国保データは各都道府県の国保連に頼めば、比較的容易に手に入れることができます。

第二に、生活習慣病対策の対象を多く含むからです。

言うまでもなく、働き盛りで若い人たちは医療機関にかかることも少なく、生活習慣病も喫緊の問題ではないことが多いと言えます。

第三に、データのフォーマットの地域差が少ないことです。

傷病名のデータ（後述します）などの数に差はあるものの、データ項目やデータ形式に大きな差がありません。どこの地方のデータでも、おおよそ同じ方法で解析することが出来ます。

これらの理由から、国保のデータを対象とします。

ただし、他の保険者のデータ形式も、極端に異なっているわけではありませんので、本稿の手法は流用することが可能です。

③注意

データの分析に取り掛かる前に、あなたが分析しようとしている国保の対象者数と、PC 環境を確認してください。

概算ですが、国保の対象者数×10 程度、レセプトは生じます。

30 万人の国保加入者がいれば、最低年間 300 万ほどのレセプトが生じると考えてください。

20000 以下のレセプトなら、EXCEL2003 で処理可能です。20000-200000 程度なら、

EXCEL2007 で処理可能です。200000-1000000 程度なら、ACCESS で処理可能です。それを超える場合は、プログラムを組むか、SAS などの統計専用ソフトに頼る必要があります。むろん、PC 環境が何年も前のものである場合、扱えるレセプト数は上記の数よりも格段に減ります（2011 年 1 月現在）。

実際のデータに本稿の内容を適用する場合は、第一にこの制約のことを考えてください。レセプト分析は容量との戦いです。30 万程度の国保加入者のレセプトは 1 年で 2GB を超えます。これは新聞縮刷版 5 年分ほどにもなるデータなのです。自分の PC 環境では無理だという場合は、データを間引いて分析する方向で考えてください。

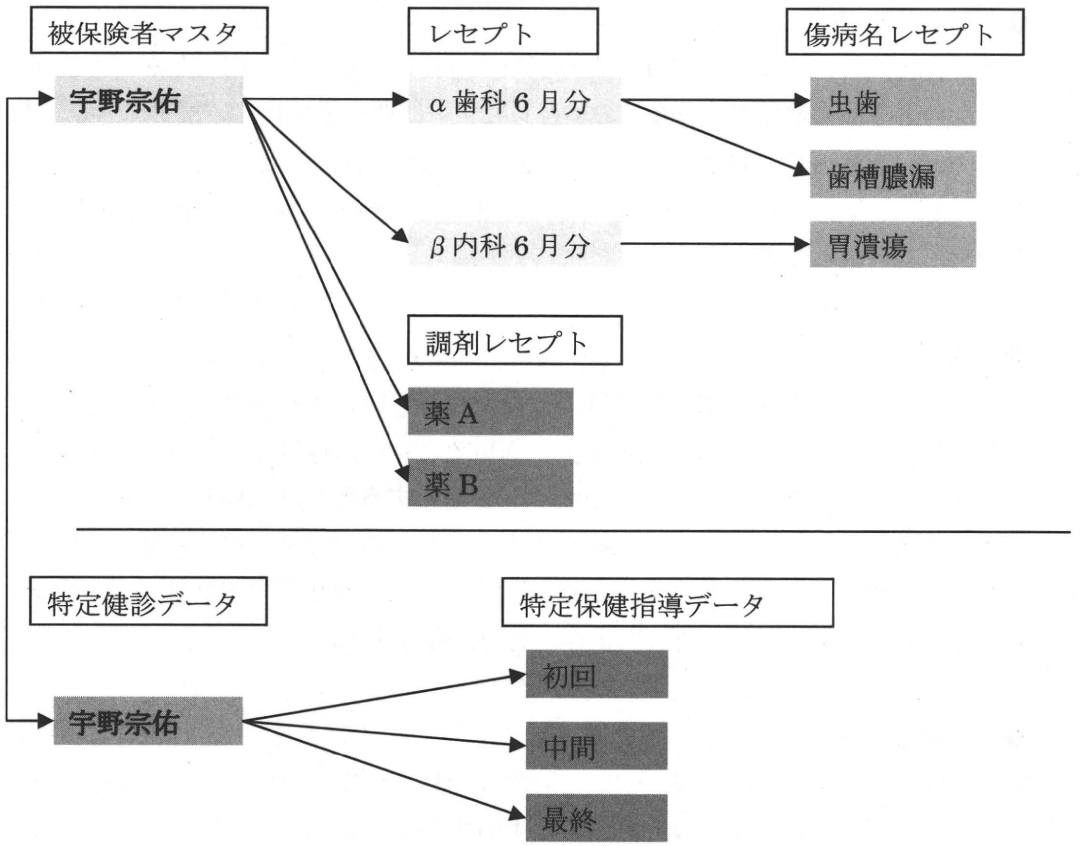
では、データの分析に取り掛かりましょう。

2.データセットの構築

①データの性質と形式

国保データは多くの場合、以下のような構成をしています。

図１　国保データ構造



●被保険者マスタ

国保の資格がある人の名簿です。当然ですが、公務員になったり、転出したり、死亡したり、国籍を失えば名簿から削除されます。逆の場合は名簿に加えられます。注意すべき点は、何カ月もたってから事後的に変更されることです。期限切れの国保を使用した場合などに、このようなことが起こります。また、転出などの場合も、リアルタイムで更新されるわけではありません。当然ながら、上記のようなこと－転入出・転職・死亡などが多いほど信頼性は低くなります。分析する地域の転入出状況などは、分析の前に抑えておいたほうがいいでしょう。

●レセプト

これに関しては多くを説明する必要はないでしょう。年齢、性別から、保険点数や入院日数、保険区分に至るまで、多くの基礎的なデータが含まれています。

レセプトデータの単位は、月にかかった医療機関数です。ある人が 4 月に歯医者に行き、接骨医に行き、精神科に行ったとすれば、この人のレセプトは 5 月に 3 枚生じることになります（データはおなじ個人番号のものが 3 行できます）。

●傷病名レセプト

地方自治体の状況によっては、このデータは存在しません。傷病名が一つしか記載されていないような自治体では、多くの場合レセプトに埋め込まれています。一方、東京 23 区などでは、病名の数に制限がないため、20 の病気に同時にかかっていれば、同じレセプト番号で 20 の病名データが存在することになります(データは同じレセプト番号のものが 20 行出来ることになります）。

●調剤レセプト

本当はこのデータを活用できればいいのですが、突合する鍵となるデータがなく、多くの自治体で利用できないのが現状です。利用できる自治体は、傷病名データの代わりにこちらを利用してください。こちらのデータを優先する理由は、傷病名データの信頼性が低いからです。周知のとおり、糖尿病の検査をしたい場合でも、重篤な糖尿病の場合でも、傷病名は「糖尿病」になります。ただし、このデータを利用するには、「インシュリン」を「糖尿病」と読み替えるような操作が必要になります。

●特定健診データ

このデータに関しては、国が定めた形式（XML 形式）に則っているため、全国共通でデータ形式が整っています。取り扱いしやすいデータです。

●特定保健指導データ

このデータも国が定めた形式（XML 形式）に則っているため、全国共通でデータ形式が整っています。特定健診データと異なるのは、介入ごとにデータが増えていくことです。たとえば、積極的支援で初回、中間、最終と三回介入したとすれば、同じ人のデータが 3 回生じることになります。データはほとんど共通しているので、多くの場合は最終データのみを用いれば事足ります。

ここまでで説明したとおり、それぞれのデータは単位が異なります。組み合わせて使用する際は、単位と組み合わせ方に気をつけて下さい。

たとえば、レセプトと特定健診データを組み合わせて分析に使うとします。その場合以下のようなパターンがあります。

表１　レセプト単位の集計

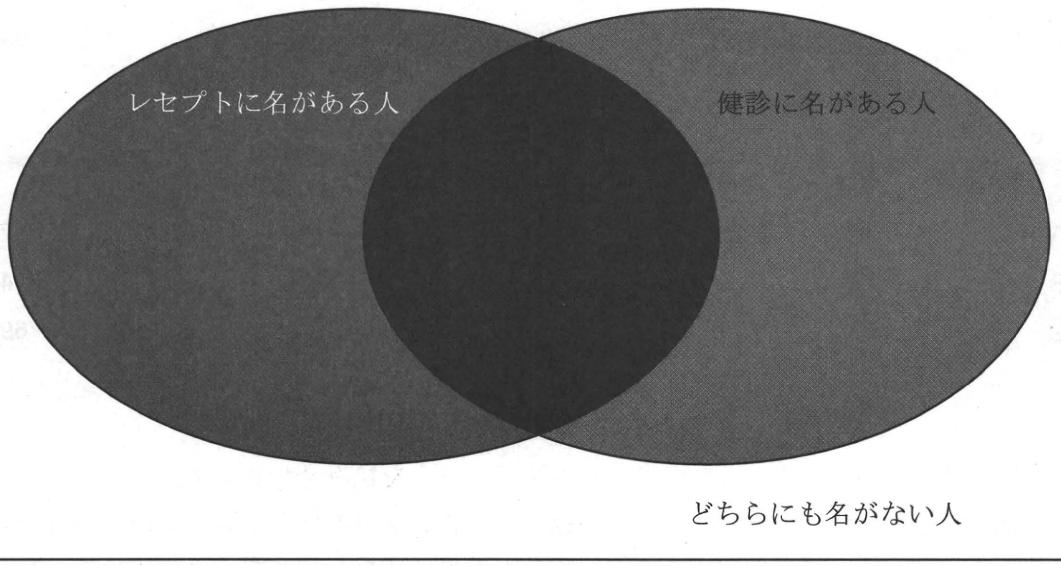
No	氏名	診療科	決定点数	特定健診実施日	体重
1	A	歯科	200	2008/10/3	55
2	A	精神科	450	2008/10/3	55
3	A	耳鼻科	300	2008/10/3	55

表２　人別の集計

氏名	特定健診実施日	体重	診療科	決定点数
A	2008/10/3	55	歯科,精神科,耳鼻科	950

さらに、多人数になると、ベン図をイメージする必要があります。

図２　国保データベン図



レセプトに名がある人（青色）を基準にデータをそろえるのか、健診に名がある人を基準にデータをそろえるのか、両方名がある人だけを抜き出すのか、あらかじめ決めておく必要があります。以下のようなデータを突合するにもいろいろなやり方があります。

健診	実施日	体重
A	39724	56
C	39726	70
D	39727	82

レセ	診療科	決定点数
A	歯科	522
B	耳鼻科	241
C	内科	695

表3　レセプトに名がある人基準

氏名	診療科	決定点数		特定健診実施日	体重
A	歯科	522		39724	56
B	耳鼻科	241			
C	内科	695		39726	70

表4　健診に名がある人基準

氏名	特定健診実施日	体重	診療科	決定点数
A	39724	56	歯科	522
C	39726	70	内科	695
D	39727	82		

表5　両方に名がある人基準（レセプト∩健診）

氏名	特定健診実施日	体重		診療科	決定点数
A	39724	56		歯科	522
C	39726	70		内科	695

表6　どちらかに名がある人基準（レセプト∪健診）

氏名	特定健診実施日	体重		診療科	決定点数
A	39724	56		歯科	522
B				耳鼻科	241
C	39726	70		内科	695
D	39727	82			

本稿では、私が一番使いよいと考える方法でデータを組み合わせます。ただし、この方法は融通がききますが、目的によっては違う組み合わせ方のほうが適している場合もあるでしょう。いくつか例を挙げますので、皆さんも目的に合わせてデータ集計の最適な方法を考えてください。

②EXCEL の操作

突合・集計作業に入る前に、簡単な EXCEL の復習を済ませておきましょう。

・＋－×÷　→　＋－＊／

・n 乗→＾n（「へ」のキー）入力例：A1^2＝A1 の 2 乗

・範囲指定　相対参照（A1:C1）

絶対参照（$A1:C$1）←始点の A 列と終点の 1 行を固定。F4 で入力

※絶対参照はコピー&ペーストしても範囲が変わらない。

・Ctrl+矢印キー　押した矢印の方向へ、データの端までジャンプ

・Shift+矢印キー　押した矢印の方向へ範囲指定

・Ctrl+Shift+矢印キー　押した矢印の方向へ、データの端まで範囲指定

最低限、相対参照と絶対参照だけは理解してから操作に入ってください。

③データの読み込み

取り扱うデータには、およそ三つの形式があります。

・CSV データ－レセプトなど

・固定長データ－傷病名など

・XML データ－健診・保健指導データなど

こう書くとややこしいですが、いずれのデータもただのテキストデータ－メモ帳で開くことができるデータです。ちなみに、メモ帳で開くと「☼景ﾆ¥・繝ﾆ m ﾆ」&－ &!&KM 鶤郎 RM ｹﾂ)ﾇ;L;L ﾇﾍ」こういう意味不明な文字列になるデータをバイナリデータといいます。話は少々脱線しますが、ひとつぐらいはテキストエディタ（高性能なメモ帳のようなもの）を用意しておくと、データのチェックの際に非常に役立ちます。お勧めは EMeditor という製品です。秀丸など他に有名なエディタはいくつもありますが、EMeditor のみが 1000 万行を超えるデータを読み込むことが出来ます。レセプトは時に 1000 万行を超えるので、事実上このソフトだけが行数を気にせず操作することが可能です。今のうちにダウンロードされることをお勧めします（もうすぐフリー版の公開が停止されるというううわさがあります）。

●CSV ファイル

CSV ファイルというのは、いわゆるカンマ区切りテキストです。

EXCEL で読み込むにせよ、ACCESS で読み込むにせよ、事前に一度テキストエディタで開いておくことをお勧めします。その理由は、「何を区切り文字に使っているか」、「文字引用符があるか」を確認するためです。

（代表的な CSV 例）

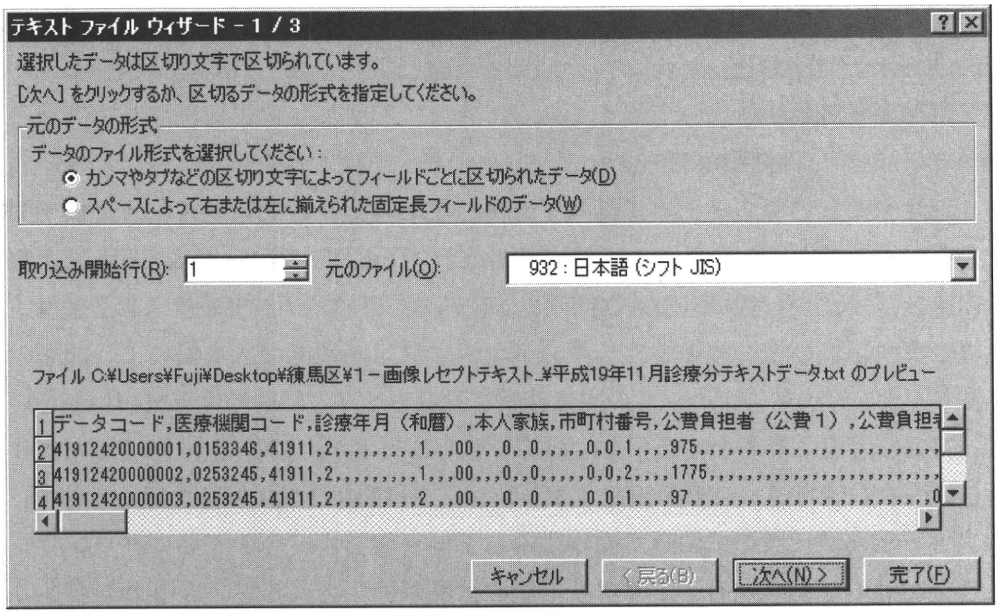
“名前”,”目”,”科”,”備考”

“イヌ”,”ネコ目”,”イヌ科”,”縄張り意識,仲間意識が強い”

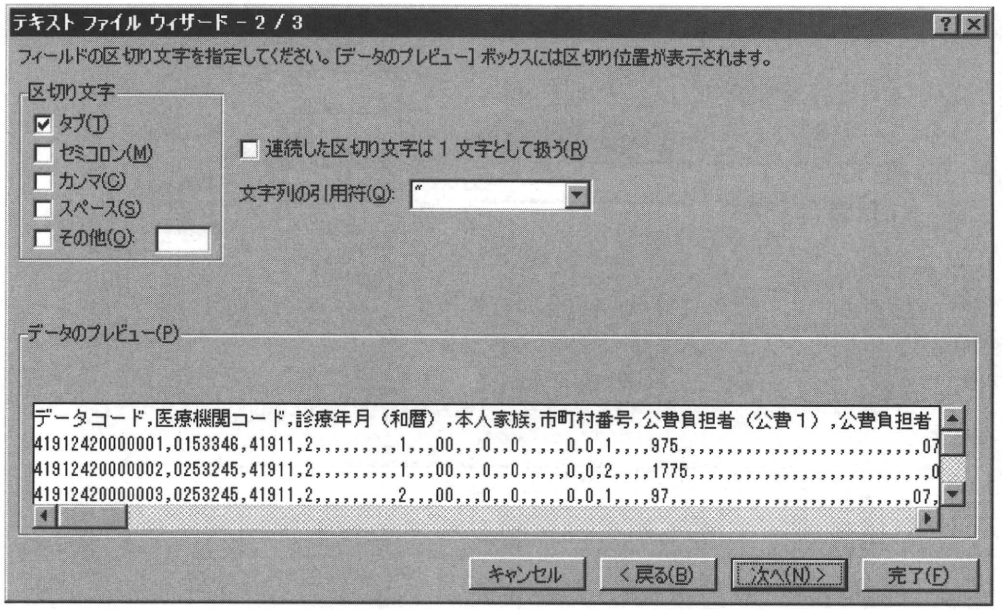
“カモノハシ”,”単孔目”,”カモノハシ科”,”毒を持つ”

この例の場合、区切り文字はカンマで、文字引用符は“です。区切り文字は特に説明の必要はないでしょう。上の例でいえば、「イヌ」と「ネコ目」をカンマで区切っています。文字引用符とは、「”と”で区切られた間はどんな文字があっても文字です」という目印のことです。なぜこのような文字が必要であるかというと、”縄張り意識,仲間意識が強い”この文中のカンマは区切り文字のカンマではない、ということを表現するためです。文字引用符がないと、EXCEL や ACCESS は、イヌの行のデータだけが 5 列、残りの行が 4 列のデータであると認識してしまい、列がずれてしまいます。

CSV の形式を確認したら、EXCEL や ACCESS で読み込んでみましょう。EXCEL の場合、とくに意識することなく読み込むことが出来るはずです。読み込んだデータがずれる場合は、たいてい上述の文字引用符が原因です。テキストエディタなどで CSV を読み込み、TXT ファイルに直して保存したうえで、再度ファイルを読み込もうとすると、テキストファイルウイザードという機能が自動的に立ち上がるはずです。

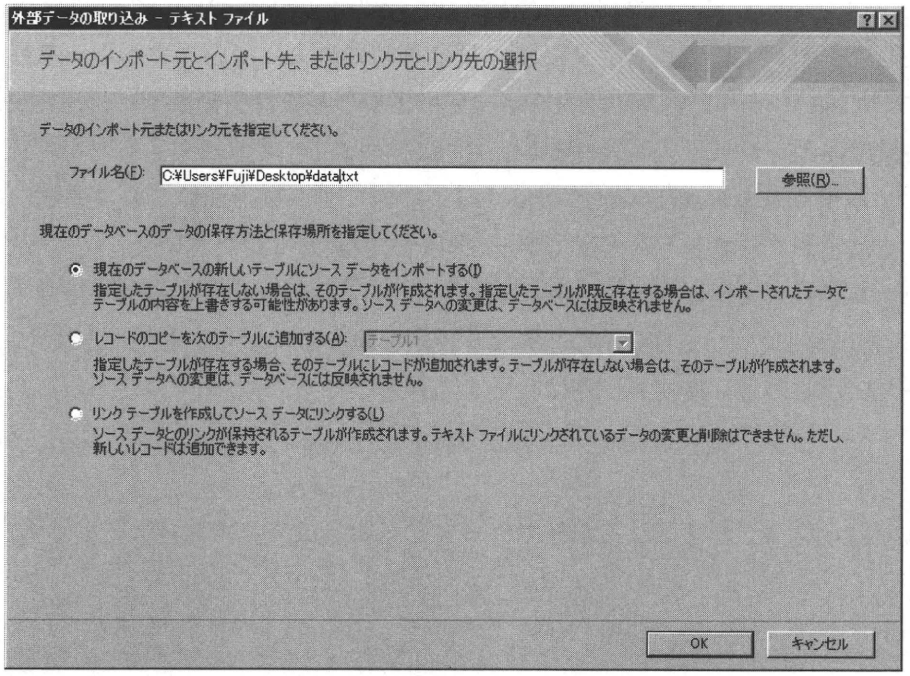


ここで、上部のラジオボタンは「カンマやタブなどの区切り文字によってフィールドごとに区切られたデータ」を選択してください。

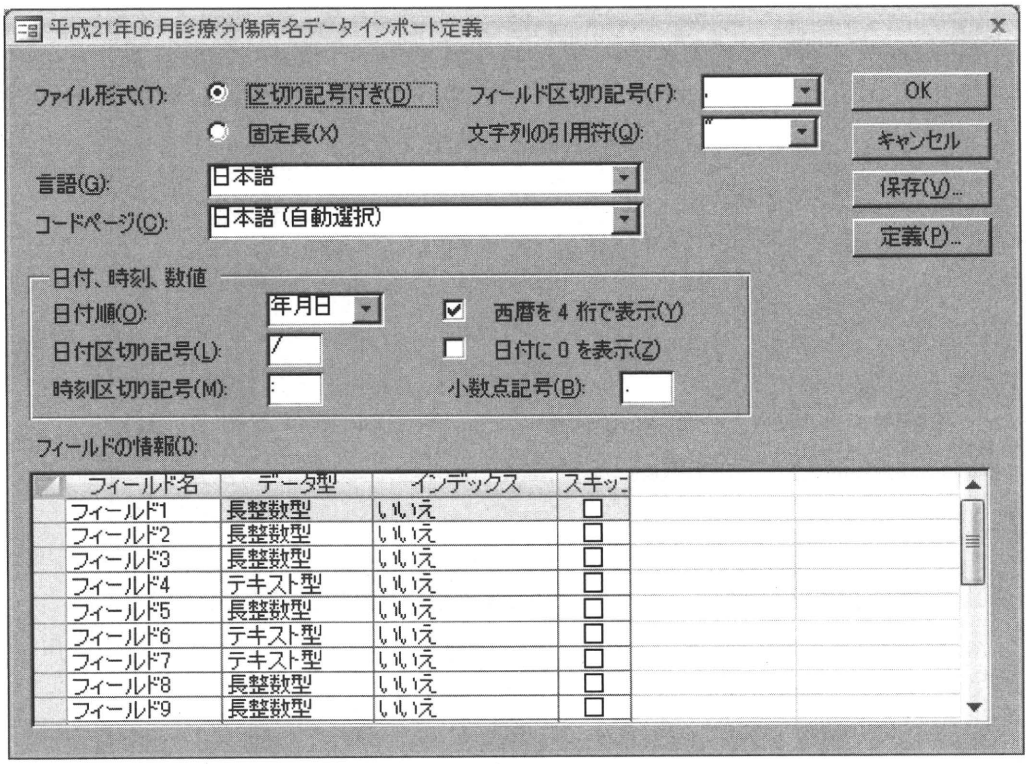


「次へ」をクリックし、この画面が表示されたら、適切な区切り文字と文字列の引用符を選択してください。これで正しく読み込めるはずです。

ACCESS で読む場合は、まず空のデータベースを作成し、「外部データ」→「インポート」からテキストデータを選択します。



下部のラジオボタンは「現在のデータベースの新しいテーブルにソースデータをインポートする」を選択しましょう。読み込むファイルを選択したら、OK を選択し、次画面で「設定」ボタンを押します。



ここで、「区切り記号付き」を選択し、各データのデータ型を選択します。ACCESS は上から数行程度のデータを読んで、データ型を自動判別します。もし、上から数行目までが数字のデータで、途中から文字データになっていると、インポートエラーが生じ文字データを読み込みません。たいていの場合、データの型は事前に決まっていますから、それを入力していってください。簡単に説明しますと、「9 桁までの整数は長整数型」、「少数があるデータは倍精度浮動小数点」、「10 桁以上の整数とすべての文字は文字列型」で読み込んでください。本当はもっと詳細に決めたほうがよいのですが、当面はこれで問題ないはずです（詳しくはインターネットなどで「データ型」で検索してください）。

右上の部分には、事前にテキストエディタ等で確認した区切り記号と文字列引用符を入力してください。最初は面倒ですが、EXCEL と異なり設定を保存しておくことができます。次回以降はスムーズになるので、最初の一回は我慢して入力してください。

●XML ファイル

XML データというのは、タグと呼ばれる記号でデータを区切ったテキストファイルです。

（XML 例）

<名前>イヌ</名前>

<目>ネコ</目>

<科>イヌ</科>

<族>イヌ</族>

＜＞で囲まれた部分をタグといいます。自分で定義することも可能ですが、健診・保健

指導ファイルの場合は事前に定められたタグが用意されています。この形式のファイルは、普通に **EXCEL** で読み込むことが出来ますが、容量が大きくなる傾向があるので、しばしば **EXCEL** の限界を超えることがあります。

●固定長ファイル

固定長ファイルというのは、テキストファイルですが、何バイト目に何のデータがあるということが決まっています。

（固定長例）

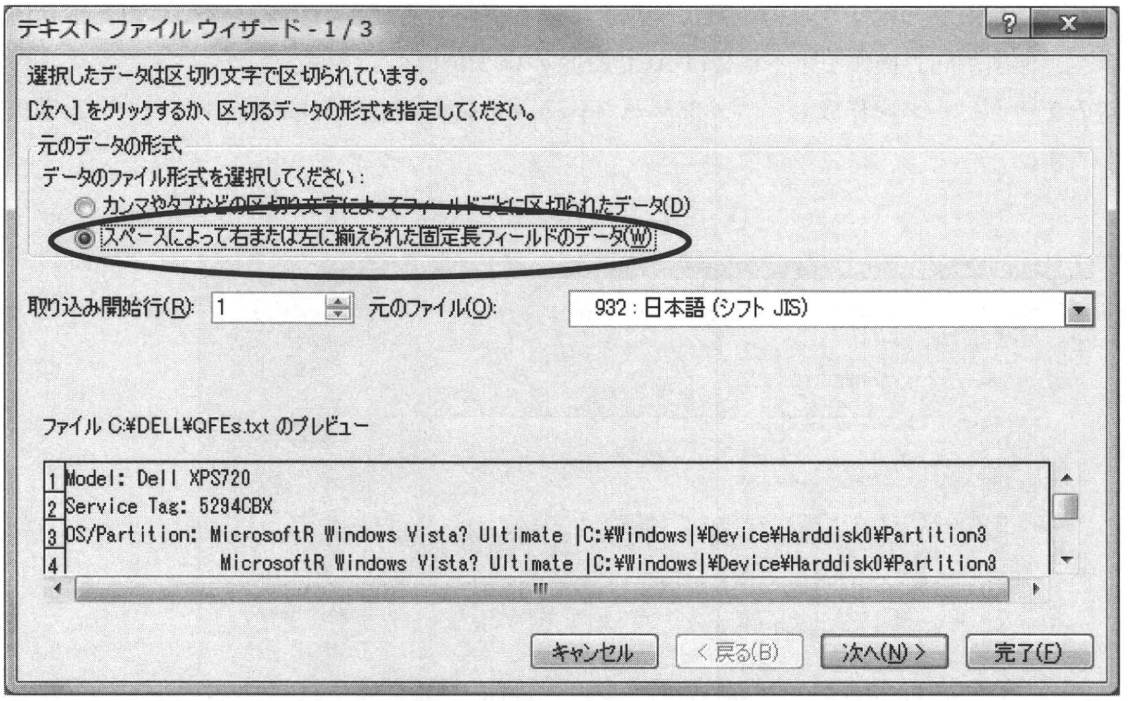
1 2 3 4 5 6 7 8 9 10111213141516171819202122 23

名前　　目　　科　　属

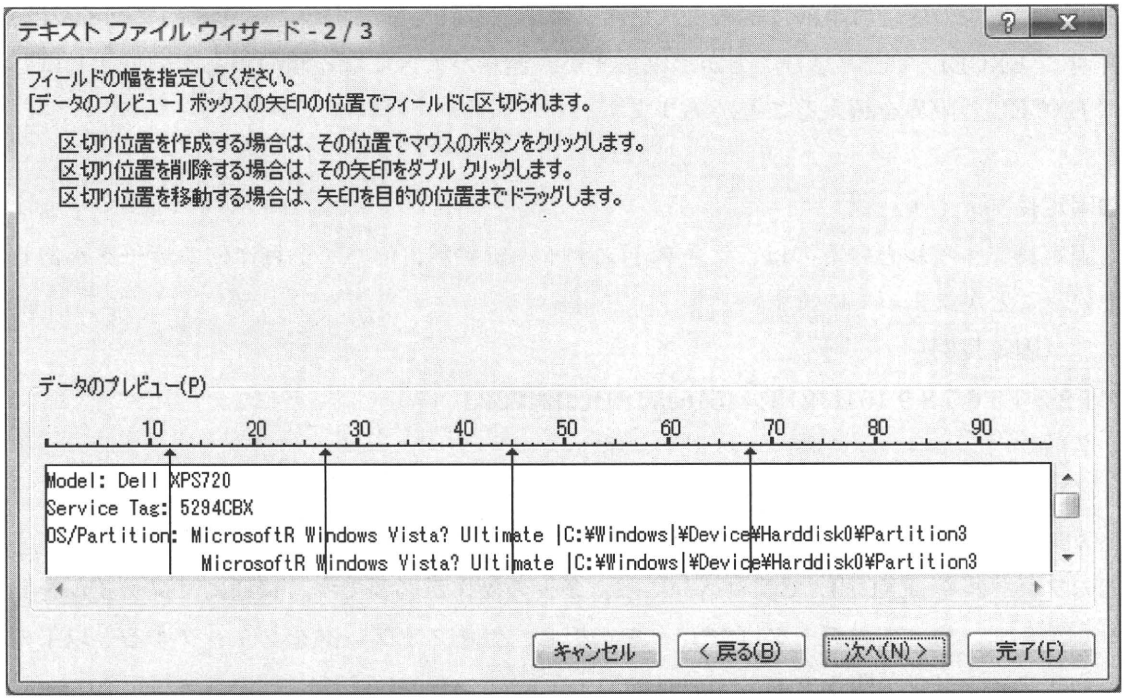
イヌ　　ネコ目　イヌ科　　イヌ属

カモノハシ　単孔目　カモノハシ科　カモノハシ属

このファイルを **EXCEL** で読み込むには、多少の操作が必要です。普通に「ファイルを開く」から、「すべてのデータ（*.*）」を選択し、テキストデータをクリックすると以下のようなウィンドウが開きます。

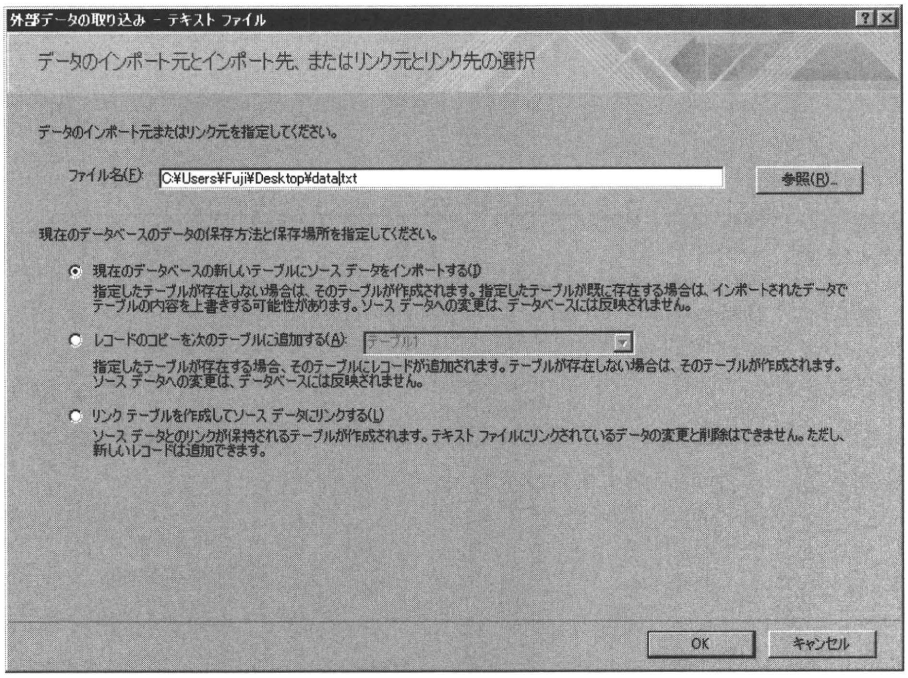


ここで、固定長を選択し、「次へ」をクリックします。



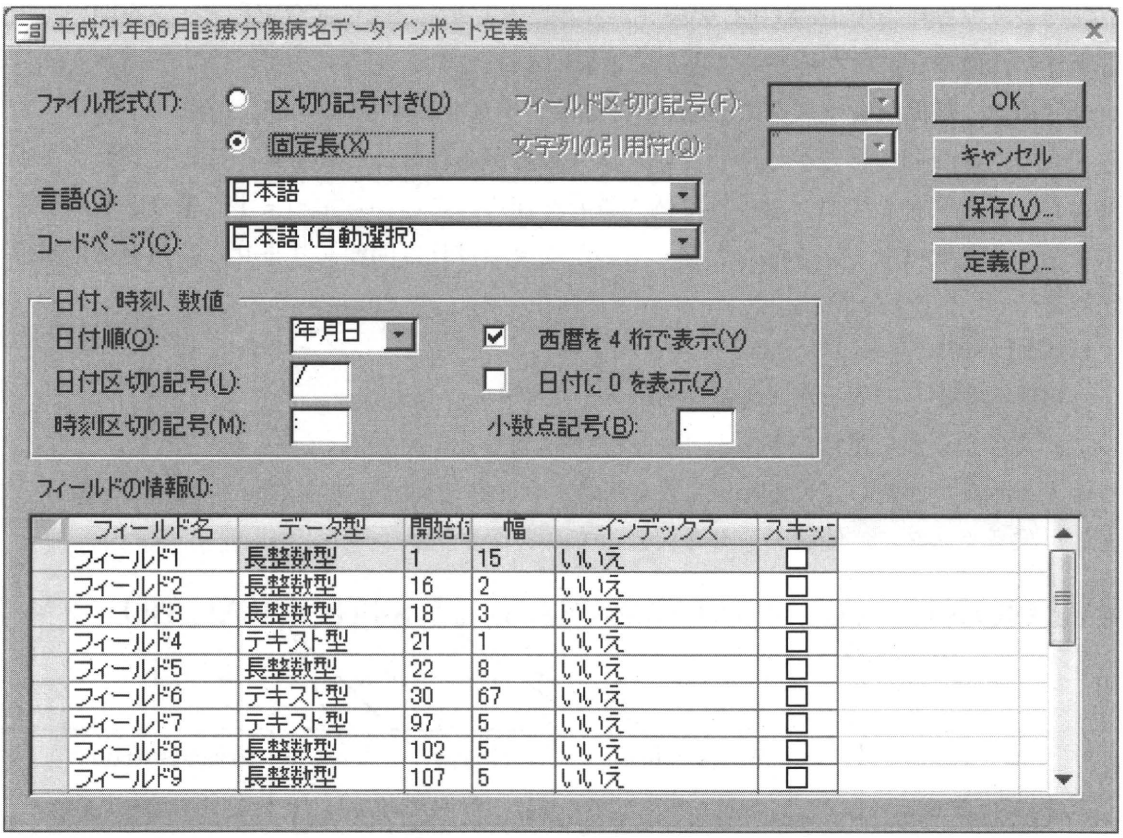
ここで、データの区切り目の位置にマウスで線を引いてください。データ項目があまりに多く、何度も同じ操作をする場合はACCESSのほうが便利です。

空のデータベースを作成し、「外部データ」→「インポート」からテキストデータを選択します。



下部のラジオボタンは「現在のデータベースの新しいテーブルにソースデータをインポ

ートする」を選択しましょう。



データ型はCSVファイルのところで説明しましたので、それを参考にしてください。
固定長ファイルの場合「n文字目からm文字目にXのデータが入る」ということが**必ず決められています(フォーマットと言います)**。それを探して入力してください。

各地方自治体においてレセプトは、似てはいますが異なるフォーマットをとっていることが多いようです（データを吐き出すシステムの問題であると思われます）。しかし、特定健診・保健指導のデータは全国共通のフォーマットです。詳細は「特定健診等データ管理システムインターフェース仕様書」を探し出してご覧ください。地方自治体の方であれば、国保年金課の方に聞けば所在が分かるはずです。

参考までに、使用するデータのフォーマットを記載しておきます。

データ	型番
被保険者マスタ	FKAB351から重複を抜いたものなど
特定健診結果	FKAC167
保健指導結果	FKAC165
判定結果（積極的・動機づけなど）	FKAC131

④データの突合

③のような操作の後、EXCELデータができたとしましょう。EXCELデータを開いてみて下さい。簡単な突合から始めてみたいと思います。

はじめに、**被保険者マスタに2008年の健診データ**を突合してみましょう

ユニークな（同じ番号が二つとない）個人番号があれば、それをキーにしてレセプトと健診データを突合します。なければ個人番号を作るところから始めます。番号と言っても数字である必要はなく、ふたつと同じものができなければ問題ありません。たとえば被保険者番号+氏名で十分です。

EXCELでは、

＝被保険者番号の入ったセル&氏名の入ったセル

で、文字や数字を連結できます。

個人番号ができたら、vlookup関数を用いてデータを突合します。

被保険者マスタに、健診データを突合してみましょう。

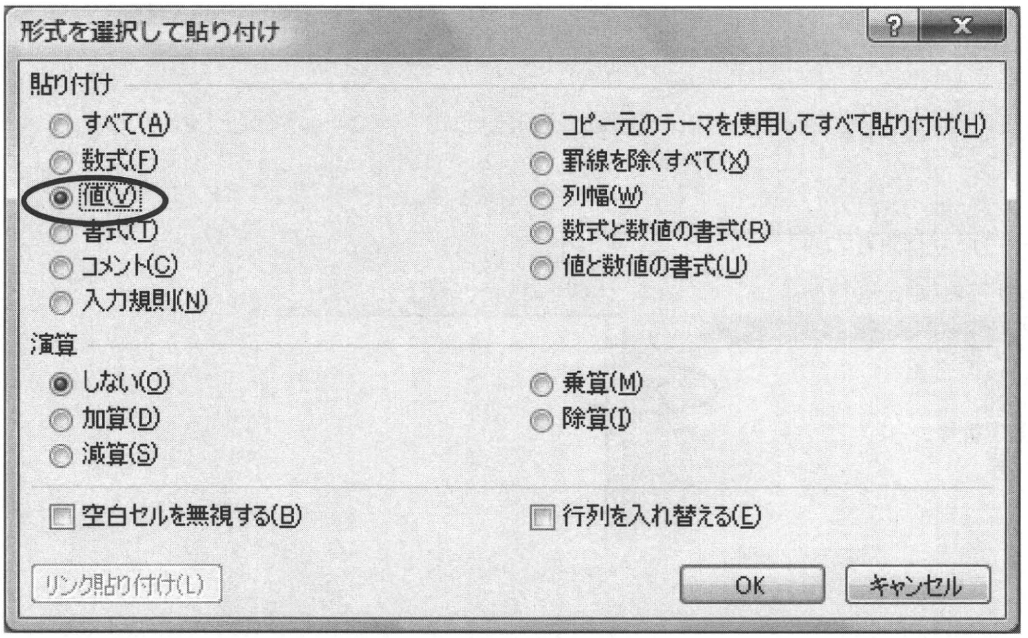
```
=VLOOKUP(A2,'2008 健診'!$A$2:$R$418,4,FALSE)
```

調べたい番号	番号簿の位置	番号簿の何列目か	近似値＝true 完全一致＝False

この式を参考に、個人番号を「調べたい番号」、健診データを「番号簿の位置」、取り出したい番号の列を指定して、入力してみてください。健診を受けた人は指定した列のデータが表示されるはずです。健診を受けていない人は「#N/A」と表示されます。うまくできたでしょうか？

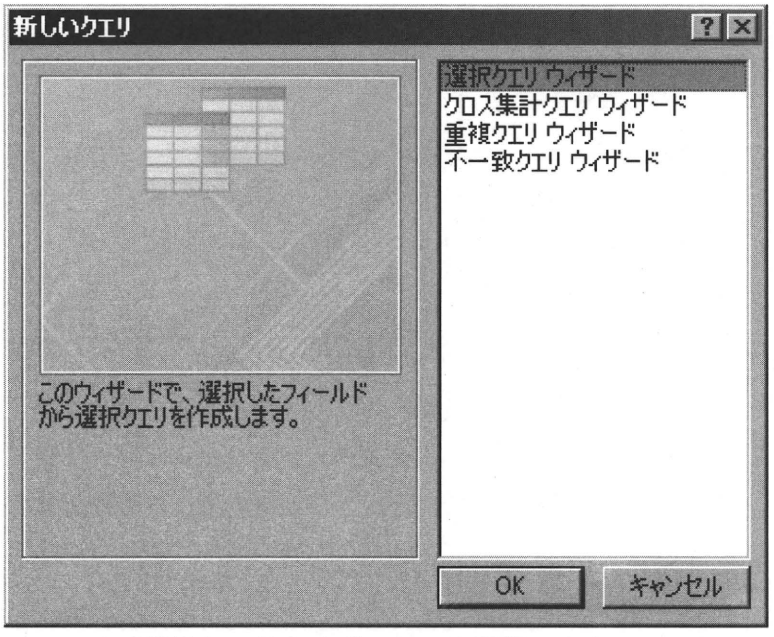
自治体の国保データを利用する際は、データの型に気を付けてください。片方が数値、片方が文字列であると、一見同じに見えても突合できません。EXCELの場合、桁数が15を超えると数値では突合できなくなりますので、両方文字列でそろえたほうがよいでしょう。

Vlookup関数が残っていると、のちの計算速度が遅くなるので、Vlookupを入力した全セルをコピーし、「形式を選択して貼り付け」から値のみを張り付けておきましょう。貼り付けたい範囲を指定したら、右クリックでメニューを呼び出し、形式を選択して貼り付けを選びます。



OK で関数なしの、値だけのデータにすることができます。

ACCESS の場合は、まずテーブルに被保険者マスタと健診データを読み込ませる必要があります。「③データの読み込み」を読んで、データの形式に応じて二つのデータを読み込ませてください。ここでは、被保険者マスタ、健診データの順に読み込ませてみましょう。レセプトのテーブルと健診データのテーブルができたら、ツールバー「作成」→「クエリウイザード」の順にクリックしてください。



ここでは「選択クエリウイザード」を選択します。