Akademie věd České republiky
Ústav teorie informace a automatizace

Academy of Sciences of the Czech Republic
Institute of Information Theory and Automation

RESEARCH REPORT

Ladislav Jirsa

Description of Data Conversion and Management in the Joint Project with the Clinic of Nuclear Medicine, Motol Hospital, Prague (Grant IGA MZ ČR 4581-3)

No. 1979 December 1999

ÚTIA AV ČR, P. O. Box 18, 182 08 Prague, Czech Republic
Fax: (+42)(2)688 4903
E-mail: utia@utia.cas.cz

Acknowledgment

This research has been partially supported by grant IGA MZ ČR 4581-3.

Contents

List of Tables

1 Introduction

This report describes ways to collect biophysical and medical data from various sources, put them together according to some requirements, perform some computations and (optionally) analyze them statistically to discover any useful relationships.

These data are results of various biophysical, biochemical, histological and other examinations of patients suffering from thyroid gland diseases and treated by ^{131}I at the Clinic of Nuclear Medicine, Faculty Hospital Motol, Prague, Czech Republic (shortly *KNM*).

The data have been processed by a software package JOD, having been run at the *KNM*, that performed computations without considering random nature or other uncertainty of these data. This processing is called *deterministic*.

The system JodNew and its algorithms, designed later, uses probabilistic approach to the uncertain data.

There are several sources of the data:

Data obtained by biophysical mesurements — "Welch-database". A software JOD provides data and results obtained during diagnostics and therapy of thyroid gland diseases. These data and results are stored in binary files `WELCH.DWW` and `*.?WW` (so-called *Welch*-format). Structure of these files is designed so that it saves space with respect to the methodology of data processing used in the system JOD. This processing is deterministic and does not respect the uncertainty of the data.

Basic database of patients at the *KNM* — "Red numbers". This file contains one record for each patient with information about his initial and final health state, results of some histological examinations etc. Each patient has a unique red number as identification. These data can be used to consider efficiency of the treatment. Name of this file is `CC.DBF`. This file has over 5 000 records.

Thyreo-globuline database. File `THYRGLO.DBF` contains information about levels of thyreo-globuline (TGL) which is an important marker of the treatment success. One record represents one TGL-examination. This file has about 40 000 records.

"Ambula" database. These data were collected manually from hospital cards for selected red numbers. This file contains some information about medical treatment (which examination was done with which result, which action was made etc.). The file is created in MS Access format and it can be easily converted to *dbf*-format `AMBULA.DBF`.

As a uniform data format, *dbf*-file was accepted. The Sequiter CodeBase libraray, that is used, can be linked with C/C++ and it is able to perform R/W operations on *dbf*-files, create files, index, sort, define relations etc. The *Welch*-data can be converted to *dbf*-format using this library.

The goals of the data conversion are mainly

- to convert data from *Welch*-format to *dbf*-format,
- to perform probabilistic-oriented computations on them,
- to connect these results to the medical data,
- to conclude advantages of the probabilistic approach for the treatment and other aspects.

In this report, important knowledge is described that is necessary for the conversion and connection.

2 Welch database description

As I heard, Welch is a name of some professor hated by everyone who knew him. I think that choice of the name for this stuff was based on a very good idea.

2.1 Data files concept

Files are organized in this way:

- They are chronologically grouped in directories with names WELCH#, where "#" means some ordinal number, e.g. WELCH15, WELCH123 etc. Directory WELCH123 is chronologically followed by WELCH124 etc. One directory contains data collected during one or two weeks.
- To save space, the data are compressed by PKZIP program. Each directory contains two files: WELCH#.ZIP with the data files and WELCH#.DIR with their listing. The symbol "#" stands for the ordinal number mentioned above.
- After decompressing, each directory contains several types of files:
  - __*.?WW, where "*" means so called "red number" and "?" means a kind of treatment. The red number is unique identification mark within the *KNM* and the kind of treatment is internal code of procedure. The most frequent numbers are 1 — basic diagnosis of a new patient, 6 — therapy without TSH (see the structure of WELCH.DWW). These files contain rough unprocessed data and they will be handled later.
  - *hexstring*.?WW, where *hexstring* is a sequence of hexadecimal characters obtained by coding of the patient's name because of absence of the red number. These files also contain data, but they are omitted from the conversion.
  - PACDIS.V*, that contain text form of so-called application protocols. These files are omitted.
  - PRPA#.OWW, where "#" stands for the *Welch*-ordinal number mentioned above, contain text form of data and results. These files are generated by the conversion program WRPAC.EXE (see below). They are omitted.
  - SEFDIS.V* that contain some (unknown) binary information. They are omitted.
  - WELCH.DWW that contain binary form of deterministic results and some additional data. These files will be handled later.

As mentioned above, for the data conversion two types of binary files will be needed: __*.?WW with the data and WELCH.DWW with deterministic results and additional data.

2.2 Structure of binary files

Now structures of binary files __*.?WW and WELCH.DWW will be described.

2.2.1 __*.?WW

This binary file has at least two records of the length of 68 bytes. The first record contains names of the measured lesions and the excretions data. The second and next records contain data of each measurement of lesions.

The fields contained in the first records are listed in Table 1, page 5. The record is created by the program with source in EXKRM.FOR.

<u>Remarks to the Table 1:</u>

1. "name" in the table header means variable name in the FORTRAN source file EXKRM.FOR.
2. Some of the lesion names can be empty, depending on how many lesions were measured (see the structure of WELCH.DWW).
3. The strings are *not* null-terminated.
4. The impulses counts are without background (deterministic subtraction because of space saving). Therefore information about excretions background is lost.

How many measurements on lesions were performed, so many 68-bytes records follow. Their structure is listed in Table 2, page 5. The records are created by the program with source in DIAGM.FOR. More information can be found in the file THYR.TXT.

<u>Remarks to the Table 2:</u>

No.	bytes	name	type	description	remark
1.	8	array `wa`	char[8]	string with name of 1^{st} lesion (usually "KRK")	
2.	8	array `wa`	char[8]	string with name of 2^{nd} lesion	
3.	8	array `wa`	char[8]	string with name of 3^{rd} lesion	
4.	8	array `wa`	char[8]	string with name of 4^{th} lesion	
5.	8	array `wa`	char[8]	string with name of 5^{th} lesion	
6.	4		float	excretion impulses 24 hours after application	10 s all below
7.	4		float	time from application to this measurement in days	
8.	4		float	excretion impulses during self-calibration	
9.	4		float	excretion impulses 48 hours after application	
10.	4		float	time from application to this measurement in days	
11.	4		float	time from application to the self-calibration in days	
	4			unused	

Table 1: Structure of the first record of _*.?WW

No.	bytes	name	type	description	remark
1.	4	array `leze`	float	number of impulses above 1^{st} lesion	100 s/10 s (diag/ther)
2.	4	array `leze`	float	number of impulses above 2^{nd} lesion	
3.	4	array `leze`	float	number of impulses above 3^{rd} lesion	
4.	4	array `leze`	float	number of impulses above 4^{th} lesion	
5.	4	array `leze`	float	number of impulses above 5^{th} lesion	
6.	4	`poz`	float	background impulses	100 s
7.	4	`def`	float	time from the application [days]	measurement time
8.	4	`eta`	float	calibration coefficient [kBq·100s/imp]	used for signal reconstruction
9.	4	`apl`	float	applied activity [kBq]	Beware! In WELCH.DWW is MBq!
	4			unused	
10.	4	`vol0*spa0*1000`	float	standard solution activity in t_0 [kBq]	
11.	4	`t0`	float	t_0 [days from 1.1.1991]	
	20			unused	

Table 2: Structure of further records of _*.?WW

1. "name" in the table header means variable name in the FORTRAN source file DIAGM.FOR.

2. Some of the lesion impulses count can be empty, depending on how many lesions were measured (see the structure of WELCH.DWW).

3. The applied activity (variable `apl`) is in kBq in contrast with Table 3, where the unit MBq is used.

4. Reconstruction of the impulses count of the standard radiator assumes that time of the standard measurement (calibration) is the same as time of the lesion measurement. Actually, lesion is measured up to two (or more) hours later.

5. The time information about calibration (standard) measurement is necessary for the standard activity determination. Unfortunately, it is not accessible as it is not stored. The time of the standard radiator measurement is approximated by the lesion measurement time, which is by few hours later (see the previous remark).

The patient is identified by the red number which is contained in the name of the file. The information about kind of treatment is contained in the file extension. Another important connective quantity is applied activity.

As all the computations done by the old system JOD are deterministic, the reconstruction of the impulses count of the standard radiator is done in the following way. Be S_0 [MBq] activity of the standard solution in time t_0 [days from 1.1.1991], T_p [days] the physical half-life of ^{131}I (8,04 days), t_m [days from 1.1.1991] the measurement time, b [impulses/100 s] count of background impulses and c [kBq·100s/impulse] the calibration coefficient. Then for the standard activity S_m in the measurement time t_m (see the previous remarks) holds

$$S_m = S_0 \exp\left(-\frac{t_m - t_0}{T_p}\ln 2\right)$$

and for count of standard impulses s per 10 s holds

$$s = \frac{1}{10}\left(\frac{1000\,S_m}{c} + b\right). \tag{1}$$

The fraction $\frac{1}{10}$ adjusts the count of standard impulses to 10 s which is the time of the standard radiator measurement (see Table 8, page 15).

In practice, it was observed that standard impulses given by the equation (1) make the standard radiator activity very "low", thus activity of lesions are very "high" and relative activities of lesions several times (cca. 8% of all cases) exceed 100%, which does not give physical meaning. On the other hand, when the fraction $\frac{1}{10}$ was omitted from (1), the results of the subsequent computations on these data seemed to be reasonable and they corresponded to those yielded by the software JodNew. The hypothesis was formulated that there is some shift in order of the calibration coefficient c when the data are stored by JOD. Nevertheless, the FORTRAN source codes have not been checked for this bug yet (20.12.1999).

2.2.2 WELCH.DWW

This binary file has as many records as many patients' actions were done in the given WELCH# directory. The record length is 214 bytes. All the records have the same structure. It is listed is in Table 3, page 6. The records are created by the program with source in WRPAC.FOR. More information can be found in the file DIAGM.TXT.

No.	bytes	name	type	description		remark
1.	2	`nc`	int	patient's number (not the red one)		*omitted*
2.	20	`jm`	char[20]	patient's name and birth year		
				[0] – [15]	patient's name	*omitted — confidential*
				[16] – [19]	birth year	
3.	8	`pac`	char[8]	patient's red number with extension		
			[0] – [5]	patient's red number		
				[6], [7]	dot, extension	
4.	2	`typ`	char[2]	type of treatment		
5.	4	`apl`	float	applied activity [MBq]		Beware! In __*.?WW is kBq!
6.	4	`hmt`	float	patient's mass [kg]		
7.	4	`hmth`	float	thyroid gland mass [g]		
8.	2	`dd`	int	day of the record creation		
9.	2	`mm`	int	month of the record creation		
10.	2	`rr`	int	year of the record creation		
11.	2	`da`	int	days after application		date of the what??
12.	2	`nl`	int	number of lesions		
13.	160	array `res`	float[40]	array with results (see below)		

Table 3: Record structure of WELCH.DWW

<u>Remarks to the table 3:</u>

1. "name" in the table header means variable name in the FORTRAN source file WRPAC.FOR.
2. Original description of the record can be found in the file DIAGM.TXT.
3. The strings are *not* null-terminated.
4. The indexing of arrays follows the C-style.
5. The applied activity (variable `apl`) is in MBq in contrast with Table 2, where the unit is kBq.

6. What is `da`? Days after application of record filling?

7. Ad `pac` — if count of the red number's digits < 6, the string is filled with underscores ('_') from left (e.g. __1234).

The structure of the results containing array `float[40]` is listed in Table 4, page 7. The structure of the array is described in the file DIAGM.TXT and it is created by program with the source file EVAL0.FOR.

Array item No.	description	remark
1.	the moment when the patient was moved to 2^{nd} stage	
2.	absolute anorganic fraction of SIK [mGy]	
3.	relative anorganic fraction of SIK [%]	
4.	the total residual activity	*omitted*
5.	effective half-life T_{ef} [days]	
6.	relative extra-thyroidal activity after 24 hours [%]	
7.	specific irradiation of blood (SIK) [mGy]	
8.	spec. irrad. of thyroid gland (SITH) [Gy]	
9]	relative activity excretion between 0–24 hours [%]	
10.	relative activity excretion between 24–48 hours [%]	
11.	AP for lesion 1	*omitted*
12.	BP for lesion 1	*omitted*
13.	T_{ef} for lesion 1	
14.	mass of thyroid gland by sonography [ml]	
15.	AP for lesion 2	*omitted*
16.	BP for lesion 2	*omitted*
17.	T_{ef} for lesion 2	
18.	mass of nodes by sonography [ml]	
19.	AP for lesion 3	*omitted*
20.	BP for lesion 3	*omitted*
21.	T_{ef} for lesion 3	
22.	spec. irrad of thyroid gland sono (SITS) [Gy]	
23.	AP for lesion 4	*omitted*
24.	BP for lesion 4	*omitted*
25.	T_{ef} for lesion 4	
26.	spec. irrad. of lymphatic nodes sono (SITN) [Gy]	
27.	AP for lesion 5	*omitted*
28.	BP for lesion 5	*omitted*
29.	T_{ef} for lesion 5	
30.	correlation coefficient for lesion 5	*omitted*
31.	residual activity of the lesion 1	*omitted*
32.	residual activity of the lesion 2	*omitted*
33.	residual activity of the lesion 3	*omitted*
34.	residual activity of the lesion 4	*omitted*
35.	residual activity of the lesion 5	*omitted*
36.	number of measurements (No. of records in __*.?WW - 1)	
37.	information about patient's red number and extension — see below	
38.	serial number of solution used for the application	*omitted*
39.	application time [days from 1.1.1991]	
40.	amount of applied radioactive solution [ml]	*omitted*

Table 4: Structure of the array with results in WELCH.DWW

<u>Remarks to the table 4:</u>

1. The ordinal number of the array item in Table 4 is *not* the array index in C-style. This notation was chosen due to compatibility with the original notation used in the old JOD system written in FORTRAN.

2. In the conversion C-program, array `float[41]` is actually used. The first item `[0]` is then empty, so the array item number in Table 4 is used as the C-array index.

3. Nos. 11, 12 etc.: AP and BP mean linear regression coefficients for radioactive decay expressed in semi-logarithmic scale. They are not used.

4. Nos. 14, 18, 22, 26: Here the correlation coefficients for lesions' 1–4 regression were stored. As this information has not been used, these positions were replaced by additional data that appeared to be stored after this database design.

5. No. 37: This position contains single digit (extension), absolute value of which is identical with the information in the 8-character array.

   - If it is positive, the patient has his/her red number.
   - If it is negative, the patient has no red number.

   Further the absolute value carries some information:

   1 basic diagnostics (newcommer),

   2 diagnostics after point 1 with TSH (correction),

   3 diagnostics directly with TSH (newcommer),

   4 therapy without TSH (newcommer),

   5 therapy with TSH (newcommer),

   6 therapy without TSH,

   7 therapy with TSH.

2.3 FORTRAN-programs for conversion binary → text

As a part of JOD system, two programs are supplied to convert binary files to legible text form:

- WRPAC.EXE that reads binary data files _*.?WW and creates text file PRPA#.OWW,
- RW.EXE that reads binary results and additional data file WELCH.DWW and converts it to a text file WELCH.CTI.

These programs originally written as a part of the old system JOD served only as revision tools to check whether the binary data files saved by JOD were read and interpreted correctly. The programs were not used for the conversion itself.

3 "Red Numbers" — CC.DBF

This file in *dbf*-format has been created at the Institute of Endocrinology. Each patient with thyriod carcinome has one record in this file.

The following table collects fields of this file and indicates which of them were used further. Some fields have been given different names following the JodNew naming convention. Type of some fields was modified for better suitability. List of the structure is in Table 5, page 9.

The fields that were taken from the mentioned file are summarized in Table 6, page 10.

Remarks to the Table 5:

1. The fields have no decimal places.

2. For each red number at most one patient exists.

3. Each existing red number has just one record.

4. These fields are updated with a new medical examination:

   (a) tnm_final

field	field name	type	width	description	new name	new type	width
1.	cislo	numeric	5	red number	cerv_cis	numeric	5
2.	prijmeni	character	25	surname	*omitted — confidential*		
3.	jmeno	character	10	first name	*omitted — confidential*		
4.	m_z	character	1	m=male, z=female	muz	numeric	1
5.	nar	numeric	4	birth year	roknaroz	numeric	4
6.	lost	character	1	a = data lost, n = not	*united with the next field*		
7.	lost_dat	numeric	4	year of data loss	ztrdata	numeric	4
8.	zemrel	character	4	year of death	roksmrt	numeric	4
9.	hacek	character	1	"hook"	*omitted — nonsense*		
10.	sono	character	1	sono. exam. y/n	*omitted — nonsense*		
11.	hacek_prav	character	1	"right hook"	*omitted — nonsense*		
12.	mct	character	1	medullar carc. ther. a/n	mct	numeric	1
13.	poznamka	character	25	remark	poznamka	memo	10
14.	jod_abl	character	1	iodine "ablation" a/n	jod_abl	numeric	1
15.	jod_th	numeric	2	preceding ^{131}I therapies	jod_th	numeric	2
16.	rok_dg	numeric	4	year of diagnosis	rokdg	numeric	4
17.	histology	character	3	histological result	histolog	character	3
18.	tnm_orig	character	3	TNM-result after reception	tnm_orig	character	3
19.	tnm_final	character	3	TNM-result after diagnosis	tnm_fin	character	3
20.	tgl_final	numeric	3	thyreo-globuline (TGL) level [μg/l]	tgl_fin	numeric	3
21.	dat_finm	character	2	month of TGL examination	*united with the next field*		
22.	dat_finr	character	2	year of TGL examination	datfin	date	8
23.	other_th	character	1	?? other therapy a/n ??	other_th	numeric	1
24.	status	character	1	patient's state	status	numeric	1
25.	typ_m1	character	4	kind of metastases	typ_m1	character	4
26.	datrecm	character	2	month of this record founding	*united with the next field*		
27.	datrecr	character	2	year of this record founding	datrec	date	8

Table 5: The Red Numbers database structure and what was used

(b) tgl_final

(c) dat_fin*

(d) status

5. The great problem is that there is no exact time information about the record update. This fact causes serious difficulties when these data are combined with other ones.

4 TGL — THYRGLO.DBF

In the file THYRGLO.DBF, the results of regular patient's checking are saved. Each checking during the patient's visit is stored in one record.

Structure of the file THYRGLO.DBF is described in the Table 7.

Remark to the Table 7: There is no exact time information about time of the TGL-examination in the Table 7. What is only available is month and year of the *record filling-in* (not the examination**!!**) in fields `datm` and `datr`. If we consider a delay of several days or weeks between the examination and the filling-in and further take into account incompleteness of the time determination, we can find out that usefullness of such an information is, politely said, problematic. It was found in many cases *impossible* to match a sequence of diagnosis, therapy and corresponding TGL-examination just because of this inapropriate form of time information.

Taking advantage of such a database, beeing kept for many years and having over 40 000 records, is thus very limited.

field	field name	type	width	description	remark
1.	cerv_cis	numeric	5	red number	
2.	roknaroz	numeric	4	birth year	
3.	muz	numeric	1	male: 0 = no, 9 = not known, otherwise yes	
4.	rokdg	numeric	4	year of diagnosis	
5.	roksmrt	numeric	4	year of death, 0 = lives, negative = dead but when?	
6.	status	numeric	1	state of the patient	
7.	ztrdata	numeric	4	year of data loss, 0=no, negative = lost but when?	
8.	other_th	numeric	1	??? other therapy ???	
9.	mct	numeric	1	medullar carcinoma therapy, 0=no, 1=yes	
10.	poznamka	memo	10	note — memo field	
11.	jod_abl	numeric	1	iodine "ablation", 0=no, 1=yes	
12.	jod_th	numeric	2	number of preceding iodine therapies	
13.	histolog	character	3	histological result	
14.	tnm_orig	character	3	TNM-result when the patient came	
15.	tnm_fin	character	3	TNM-result after patient's diagnosis	
16.	typ_m1	character	4	kind of metastases if appear in TNM-result	
17.	tgl_fin	numeric	3	thyreo-globuline level [μg/l]	
18.	datfin	date	8	date of thyreo-globuline examination	
19.	datrec	date	8	date of this record founding	

Table 6: Structure of data that were taken from Red Numbers

field	field name	type	width	decimals	description	remark
1.	cislo	numeric	5		red number of the patient	
2.	datm	character	2		month of the record filling	
3.	datr	character	2		year of the record filling	
4.	tgl	numeric	3		TGL-level in blood [μg/l]	
5.	suprese	logical	1		external hormones income .f./.t.	
6.	j131mbq	numeric	9	3	applied therapeutic activity [MBq]	
7.	akumulace	numeric	5	1	relative activity [%] in thyr. after 4 hrs.	
8.	thyr	logical	1		thyreotoxicosis .f./.t.	
9.	tumor	logical	1		tumor .f./.t.	
10.	n	logical	1		nodes affected .f./.t.	
11.	m	logical	1		metastases .f./.t.	

Table 7: Structure of the TGL-database file

5 Ambula — AMBULA.DBF

The file AMBULA.DBF was created as "retrospective work" by manual typing the data from patients' cards. This work was very exhausting and monotonous. About 1 500 records were typed.

A record of AMBULA.DBF has over 150 fields. It contains information about a patient's surgery, various treatment methods, medical treatment, examination (biochemical, X-ray, scintigraphic etc.), plan for the next treatment etc.

The structure of the table is described below. It is available only in Czech.

```
Struktura tabulky ambula.dbf (ambulantní kontroly)

  č. název     typ délka des. komentář
  ===============================================================
  1. id         n    4        pořadí -- identifikační číslo záznamu
  2. cerv_cis   n    5        červené číslo -- identifikační číslo pacienta
  3. dat_kontr  d    8        datum kontroly
  4. delta_t    n    2        počet měsíců od minulé akce
  5. minule     n    1        minulá akce:
                                0 -- hospitalizace na KNM
                                1 -- ambulantní kontrola na KNM
```

```
                                   2 -- jiné -- uvést (viz dále)
  6. co_minule  c   40           pokud minule=2, popis minulé akce
  7. mezidobi   n    1           stav od minulé akce:
                                   0 -- v pořádku
                                   1 -- nesouvisející onemocnění
                                   2 -- zhoršení stavu -- uvést (viz dále)
  8. zhors_stav c   40           pokud mezidobi=2, popis zhoršení stavu
  9. operace    l    1           pokud .f., nic.
                                 pokud .t., vyplní se tato následující pole op_*:
----
 10. op_dat     d    8           datum operace
 11. op_prac    c   20           pracoviště
 12. op_mudr    c   20           operatér
 13. op_typ     n    1           typ výkonu
                                   0 -- dokončení do TTE l. utr.
                                   1 -- extirpace uzlin(y) syndromu
                                   2 -- extirpace zbytku tumoru
                                   3 -- jiné -- uvést (viz dále)
 14. ot_jine    c   40           pokud op_typ=3, popis
 15. op_velik   n    3           velikost nádoru, jednotky viz níže
 16. ov_jednot  c    3           jednotky
 17. op_hist    n    1           histologický nález
                                   0 -- benigní
                                   1 -- maligní -- viz dále
                                   2 -- nejasný
***                                3 -- neznámý
 18. oh_maligni n    1           pokud op_hist=1
                                   0 -- papilární (p)
                                   1 -- folikulární (f)
                                   2 -- smíšený (pf)
                                   3 -- jiný -- viz níže
 19. om_jiny    c   40           pokud oh_maligni=3, popis
 20. op_komplik n    1           komplikace při operaci
                                   0 -- nic
                                   1 -- paréza n. lar. rec l. dx
                                   2 -- paréza n. lar. rec l. sin.
                                   3 -- paréza n. lar. rec bil.
                                   4 -- hypoparatyreóza
                                   5 -- jiná, viz níže
 21. ok_jine    c   40           pokud op_komplik=5, popis
 22. op_pozn    c   40           poznámka
----
 23. ex_ozar    l    1           zevní ozáření,
                                   pokud .f., nic
                                   pokud .t., vyplnit tato pole eo_*:
----
 24. eo_davka   n    5   2       celková dávka [Gy]
 25. eo_pocfr   n    2           počet frakcí
 26. eo_pole    c   40           kam se zářilo
 27. eo_typ     n    1           typ ozáření
                                   1 -- Co (kobalt)
                                   2 -- betatron
                                   3 -- jiné, viz níže
 28. et_jine    c   40           pokud eo_typ=3, popis
 29. eo_prac    c   20           pracoviště
 30. eo_datfin  d    8           datum dokončení
 31. eo_pozn    c   40           poznámka
```

```
----
 32. neurochir  l    1        neurochirurgická operace
                                .f. -- nic
                                .t. -- vyplnit pole no_*
----
 33. no_datum   d    8        datum operace
 34. no_prac    c   20        pracoviště
 35. no_typ     c   40        typ výkonu
 36. no_pozn    c   40        poznámka
----
 37. subjekt    l    1        subjektivní stav: .t. OK
                                                 .f. vyplnit pole sb_*
----
 38. sb_unav    l    1        únava
 39. sb_hub     l    1        hubnutí
 40. sb_palp    l    1        palpitace
 41. sb_poc     l    1        pocení
 42. sb_chrap   l    1        chrapot
 43. sb_dus     l    1        dušnost
 44. sb_jine    c   40        jiné
----
 45. suprese1   n    1          0 -- nic
                                1 -- Euthyrox50
                                2 -- Euthyrox100
                                3 -- Euthyrox150
 46  su1_tbl    n    1        počet tablet
 **. su1_stop   n    2        počet týdnů od vysazení suprese
 47. suprese2   n    1          0 -- nic
                                1 -- Tertroxin
 49. su2_tbl    n    1        počet tablet
 **. su2_stop   n    2        počet týdnů od vysazení suprese
 47. suprese_j  c   20        jiná suprese
 49. suj_tbl    n    1        počet tablet
 **. suj_stop   n    2        počet týdnů od vysazení suprese

 49. tolerance  l    1        tolerance
                                .t. -- dobrá
                                .f. -- vyplnit pole to_slovne
 50. to_slovne  c   40        popis tolerance

objektivní nález
----------------
 51. nalez      l    1        objektivní nález,
                               .f. -- přiměřený věku a nesouvisející onemocnění
                               .t. -- vyplnit pole na_slovne
 52. na_slovne  c   40        popis nálezu
 53. krk_palp   l    1        palpace na krku
                               .f. -- negativní, nic
                               .t. -- pozitivní, vyplnit pole re_*
----
 54. re_pocet   n    2        počet rezistencí
 55. re_velik   n    3        velikost rezistencí
 56. re_jedn    c    2        jednotky (cm, ml)
 57. re_lokal   c   40        lokalizace
----
 58. tetan      l    1        tetanické projevy, .f. -- negativní, nic,
                                                   .t. -- vyplnit pole te_*
```

```
 59. te_tr      n    3        Trousseau -- čas, kdy křeč [s]
 60. te_chv     l    1        Chvostek, .f. -- negativní, .t. -- pozitivní
 61. fo_zretel  l    1        fonace zřetelná
 62. fo_ticha   l    1        fonace tichá
 63. fo_chrap   l    1        fonace chrapot
 64. fo_afo     l    1        fonace afonace
 65. ras        n    3        reflex Achillovy šlachy [ms]
 66. hmpac      n    3        hmotnost pacienta [kg]
 67. vyska      n    3        výška pacienta [cm]

laboratoř
---------
 68. sstsh      n    6    2   ssTSH [mIU/l]
 69. tgl        n    6    1   tyreoglobulin [$\mu$g/l]
 70. a_tgl      c    3        anti-tgl (takové divné křížky a plusíky)
 71. vys_jina   l    1        byla provedena jiná vyšetření
                               .f. -- nic
                               .t. -- ??
 72. jv_slovne  c   40        slovní popis jiného vyšetření

USG
---
 73. usg        l    1        USG, .f. -- nebylo, nic, .t. -- bylo, vyplnit us_*
 74. us_mudr    c   20        jméno vyšetřujícího
 75. us_nalez   l    1        nález .f. -- negativní, nic, .t. pozitivní, usp_*
 76. usp_zb_nr  n    2        počet zbytků tkáně vpravo
***. usp_zb_or  n    4        objem zbytků vpravo [ml]
 76. usp_zb_nl  n    2        počet zbytků tkáně vlevo
***. usp_zb_ol  n    4        objem zbytků vlevo [ml]
 78. usp_zbyt_l c   40        lokalizace zbytků
 79. usp_uz     n    2        počet uzlin
 80. usp_uz_o   n    3        objem [ml]
 81. usp_uz_l   c   40        lokalizace uzlin
 82. usp_jine   c   40        jiný nález
 83. usp_pozn   c   40        poznámka

další
-----
 84. dalsivys   l    1        další vyšetření, .f. nebyla, nic, .t. byla, dv_*
 85. dv_typ     c   40        typ dalšího vyšetření
 86. dv_mudr    c   20        jméno vyšetřujícího
 87. dv_vysl    c   40        výsledek dalšího vyšetření
 88. dv_pozn    c   40        poznámka k dalšímu vyšetření

přinesené výsledky
------------------
 89. usg_krk    l    1        USG krku, .f. -- nebylo, .t. -- usk_*
 90. usk_datum  d    8        datum vyšetření
 91. usk_prac   c   20        pracoviště
 92. usk_mudr   c   20        vyšetřující
 93. usk_vysl   c   40        výsledek
 94. usk_pozn   c   40        poznámka

 95. fnab       l    1        FNAB (fine needle aspiration biopsy)
                               .f. -- nebylo, .t. -- fna_*
 96. fna_datum  d    8        datum vyšetření
 97. fna_prac   c   20        pracoviště
```

```
 98. fna_mudr   c   20        vyšetřující
 99. fna_vysl   n    1        výsledek
                                0 -- benigní
                                1 -- maligní
                                2 -- nejasný
100. fna_pozn   c   40        poznámka

101. orl        l    1        ORL, .f. -- nebylo, .t. -- orl_*
102. orl_datum  d    8        datum vyšetření
103. orl_prac   c   20        pracoviště
104. orl_vysl   l    1        výsledek, .f. -- v normě, .t. -- viz níže
105. orl_popis  c   40        pokud orl_vysl=.t., popis
106. orl_pozn   c   40        poznámka

107. rtg_ct     l    1        RTG/CT, .f. -- nebylo, .t. -- rct_*
108. rct_typ    c   40        typ vyšetření
109. rct_datum  d    8        datum vyšetření
110. rct_prac   c   20        pracoviště
111. rct_vysl   l    1        výsledek, .f. -- v normě, .t. -- viz níže
112. rct_popis  c   40        pokud rct_vysl=.t., popis
113. rct_pozn   c   40        poznámka

114. sci_skel   l    1        scintigrafie skeletu .f. -- nebylo, .t. -- scs_*
115. scs_datum  d    8        datum vyšetření
116. scs_prac   c   20        pracoviště
117. scs_vysl   l    1        výsledek, .f. -- negativní, .t. -- viz níže
118. scs_popis  c   40        pokud scs_vysl=.t., popis
119. scs_pozn   c   40        poznámka

120. gynekol    l    1        gynekologie .f. -- nebyla, .t. -- gyn_*
121. gyn_datum  d    8        datum vyšetření
122. gyn_prac   c   20        pracoviště
123. gyn_vysl   l    1        výsledek, .f. -- přiměřený věku, .t. -- viz níže
124. gyn_popis  c   40        pokud gyn_vysl=.t., popis
125. gyn_pozn   c   40        poznámka

126. ostatni    l    1        ostatní vyš. .f. -- nebylo, .t. -- ost_*
127. ost_typ    c   40        typ vyšetření
128. ost_datum  d    8        datum vyšetření
129. ost_prac   c   20        pracoviště
130. ost_vysl   l    1        výsledek, .f. -- OK, .t. -- viz níže
131. ost_popis  c   40        pokud ost_vysl=.t., popis
132. ost_pozn   c   40        poznámka

závěr
-----
133. zaver      l    1        .t. -- "dobry"  ???, .f. -- za_slovy
134. za_slovy   c   40        popis závěru

plán
----
135. amb_mesicu n    2        za kolik měsíců ambulantní kontrola
136. hospit     l    1        hospitalizace .f. -- neplánována, .t. -- ho_*
137. ho_mes     n    2        za kolik měsíců plánována hospitalizace
138. ho_let     n    2        za kolik let plánována hospitalizace
139. ho_pozn    c   40        poznámka k hospitalizaci
140. jiny_plan  c   40        jiný plán -- popis
```

```
kód TNM
-------
141. tnm_t      c    1        T: x, 0, 1, 2, 3, 4
142. tnm_t_ab   c    1        T*: a, b
143. tnm_n      c    1        N: 0, 1
144. tnm_n_ab   c    1        N*: a, b
145. tnm_m      c    1        M: 0, 1
146. tnm_m_s    l    1        M*: s
147. tnm_m_p    l    1        M*: p
148. tnm_m_ss   l    1        M*: ss
149. tnm_m_pp   l    1        M*: pp
150. tnm_jine   c   20        TNM: jiná než výše
```

6 Ways of the data conversion

In this section, ways of various conversions of data and connections of *dbf*-tables are described.

6.1 Conversion WELCH $\longrightarrow$ DBF

The Welch-data were forst converted to the *dbf*-format because of data compatibility.

For the Welch-data, similar idea of dbf-files was accepted. Two files were generated:

1. "Master" file named _WDBF.DBF that corresponds to the file WELCH.DWW and contains information about the patient, various masses, excretions, effective half-lives, irradiations etc. Each record of this file contains information about one application of ^{131}I, both diagnostic or therapeutic.

2. "Slave" file named _WDATA.DBF containing "raw" data (standard activities, impulses corresponding to the standard radiator and to lesions measured). Each record contains information about one patient's measurement.

The difference is that each file collects information about all the patients, i.e. all the information is contained in two files.

What is missing in the original binary Welch-files is

1. Background for excretions. As the measurement methodology and device for excretions is different from lesions, background impulses in lesions measurement cannot be used for excretions. Therefore information about background during measurement of excretions is lost.

2. Duration of various mesurements (background, standard, lesions, excretions). This information can be easily retrieved as the measurement methodologies remain unchanged for many years. The durations are specified in Tables 8 and 9. For example, if a lesion is measured in therapy, background is measured for 100 s, standard for 10 s, calibration coefficient is recalculated to 100 s (a feedback for the staff) and signal of the lesion is measured for 10 s.

3. Information about the standard radiator activity and corresponding impulses. If the assumption is accepted that standard and lesion measurement times are the same, this information can be retrieved using (1), page 6 (actually, lesions are measured up to several hours later than the standard).

	lesions [s]	excretions [s]
background	100	10
standard on background	10	10
recalculation coefficient for c	100	10

Table 8: Durations of measurement for standard radiators

	lesions [s]		excretions [s]
	diagnosis	therapy	
signal on background	100	10	10

Table 9: Durations of measurement for various signals

Structure of the file _WDBF.DBF was extended by Bayesian estimates of various quantities. The Bayesian estimate of a quantity x consists of two parts: point estimate $\hat{x}$ and its uncertainty Δx, so that the estimate can be written in the form $x = \hat{x} \pm \Delta x$. The uncertainty represents 95%-credibility interval. Field names with point estimates end uniformly with _o ("odhad", e.g. `tef_o`) and field names with uncertainties end uniformly with _ch ("chyba", e.g. `tef_ch`).

The conversion is done by program w2dbf.cpp using Sequiter CodeBase 4.5 library. Structures of _WDBF.DBF and _WDATA.DBF are described in strucpw.h which is included into w2dbf.cpp.

Pointer to array of fields in _WDBF.DBF is `strucw`, pointer to array of fields in _WDATA.DBF is `strucwd`. These arrays are defined in strucpw.h.

6.2 Connection to __*.?WW data files and Bayesian computations

Both the files _WDBF.DBF and _WDATA.DBF are indexed to be connected together, because data from _WDATA.DBF are read for the computations and the results are stored in the corresponding file of _WDBF.DBF.

The file _WDBF.DBF is sorted according to red number and date of ^{131}I application. The file _WDATA.DBF is sorted according to red number, date of ^{131}I application, date of measurement and hour of measurement. The summary is in the Table 10. In this way, the right data in _WDATA.DBF are connected

file	compound index pointer	tag name	sort order
_WDBF.DBF	`wtag`	`"ccdat"`	`cerv_cis, datapl`
_WDATA.DBF	`wdtag`	`"wrelat"`	`cerv_cis, datapl, datmer, hodmer`

Table 10: Sorting of Welch-data files in *dbf*-format

to the right record in _WDBF.DBF.

This connection is done by the program bayewlch.cpp which performs Bayesian computations on the data.

6.3 Connection of Welch and Red Numbers

In this task, data from CC.DBF were connected to _WDBF.DBF. As it was mentioned, time information in *dbf*-files created by physicians at the *KNM* is very vague (see sections 3 and 4). This fact complicated the connection very much and, as the result, the "time intersection" of CC.DBF and _WDBF.DBF had only less than 300 records. Some "principle of uncertainty" was observed: the more time tolerance, the more records in intersection, a vice versa. Therefore this way of research was rejected.

Program to connect Welch and Red Numbers is called prwelch.cpp. Structure of the desired file is described in array pointed by `strucpw` and structure of compound index is described in array pointed by `ptag`. Both these arrays are defined in strucpw.h.

6.4 Connection of Welch and Ambula

The attempt to transform "paper data" to electronic form was not too successful. Although the file AMBULA.DBF had about 1 500 records, it was very sparse, many fields were filled in a non-consistent way and a lot of information was missing without any chance of retrieval. In this situation, this connection of Welch and Ambula was tried:

1. A destination *dbf*-file was created that contained all fields from _WDBF.DBF and AMBULA.DBF. The duplicated fields were of course removed.

2. For each red number found both in Welch and Ambula, the corresponding records were combined "each with each" to the destination file.

3. Some criteria to eliminate records that made no sense were proposed together with attempts of careful statistical analysis.

The analyses brought nothing significant.

But it was found that the most used quantity taken from Ambula is TGL-level in blood, which is an important marker of therapy success. Therefore it was decided to omit other "paper" data and to focus on TGL-database having been maintained in a file THYRGLO.DBF for several years.

Program to connect Welch and Ambula is called ambwelch.cpp. Structure of the destination file is obtained by merging the structures of both the input files. Pointer to compound index `atag` is defined in ambwelch.cpp.

6.5 Connection of Welch, TGL and Red Numbers

In this task, the files _WDBF.DBF, THYRGLO.DBF and CC.DBF were connected. The connection was done in this way:

1. Find a sequence of three records in _WDBF.DBF: diagnostics, immediate subsequent therapy (few days later) and next diagnosis for the same patient.

2. Find a sequence of records in THYRGLO.DBF: first, second and third TGL-examination (date + result) *before* therapy, first, second and third TGL-examination *after* therapy and the last avilable TGL-examination.

3. Find some information in the corresponding record of CC.DBF.

4. Put all this information together into one record of the destination file.

Program to connect Welch, TGL and Red Numbers is called tglwelch.cpp. Structure of the destination file is described in array `stglw`, index for THYRGLO.DBF is defined in `ttag`, index for the destination file is defined in `twtag`. All these arrays are defined in the source file tglwelch.cpp.

Because of problems with time information in THYRGLO.DBF (see section 4), it was often impossible to resolve if some TGL-examination was done *before* or *after* therapy. This uncertainty lead to decision to choose a different marker than TGL. As an alternative marker, maximum relative activity of thyroid gland in diagnostics after therapy (field `na1rm_o`) was chosen. With this choice, some reasonable results were given by statistical analysis.

6.6 Source C-files used

Watcom C++ 10.5 was used as a C-programming tool.

The Table 11 gives list of source files of programs used in data manipulations. Table 12 gives information about the programs building.

Several simple data manipulations and computations were done in FoxPro environment manually or using programs in FoxPro (*.PRG). These programs are not mentioned here.

7 JodNew — a new source of Welch-data

Since September 1999, old JOD has been replaced by JodNew at the *KNM*. The software system JodNew written in MS Visual FoxPro provides data in *dbf*-format and uses computational programs written in C++ to perform Bayesian computations on data.

The files with data and results in a legible "old" *dbf*-format are generated by a class `sber_dat` which is a part of the system JodNew. The class generates two kinds of files:

1. WJNRMMYY.DBF, where MM is month and YY is year without century (it started four months before 2000, so it may not matter too much). R means "root". This file is by its content similar to WELCH.DWW or _WDBF.DBF. Structure and field names follow the conventions of _WDBF.DBF.

C-files:

name	note
ambula.cpp	unused
ambwelch.cpp	connection of Ambula + Welch
baywelch.cpp	Bayesian computations on Welch data
date.cpp	date manipulations
deldup.cpp	deleting duplicated records in *.DBF
hex2dec.cpp	conversion hexadecimal ⟶ decimal
index.cpp	indexing the *dbf*-file
pac.cpp	just to display content of __*.?WW
prwelch.cpp	connection of Red Numbers + Welch
putext.cpp	function to manipulate with file extensions
tglwelch.cpp	connection of Red Numbers, TGL and Welch
w2dbf.cpp	conversion binary Welch ⟶ *dbf*
welch.cpp	just to display content of WELCH.DWW

H-files:

name	note
ambula.h	structure of AMBULA.DBF — unused
baywelch.h	structures and classes definitions etc.
codebase.h	includes d4all.h only once
putext.h	declaration of `putExtension`
strucpw.h	structures of various *dbf*-files and their indexes
welch.h	some definitions, declarations etc.

Table 11: List of the source files for data manipulations programs

2. WJNLMMYY.DBF, where MM is month and YY is year without century. L means "lesion". This file is by its content similar to __*.?WW or _WDATA.DBF. Structure and field names follow the conventions of _WDATA.DBF.

Data and results of patients applied by ^{131}I in the given month are stored in these files. Therefore they must be kept in the directory of JodNew until it is *sure* that no patient applied in the previous month is measured any more.

22.12.1999, w2dbf.tex

Task: Conversion binary Welch ⟶ DBF

exe-file	makefile	link file	modules	included	included
w2dbf.exe	—	w2dbf.lnk	w2dbf.cpp	strucpw.h	—
				welch.h	codebase.h
				date.cpp	—
				deldup.cpp	welch.h
					putext.cpp
			hex2dec.cpp	—	—

Task: Bayesian computation on Welch *dbf*-files

exe-file	makefile	link file	modules	included	included
bayw_w.exe	bayw_w.wmk	bayw_w.lnk	baywelch.cpp	strucpw.h	—
				putext.h	—
				baywelch.h	—
			putext.cpp	welch.h	codebase.h
			date.cpp	—	—

Task: Connection of Welch and Red Numbers

exe-file	makefile	link file	modules	included	included
prwelch.exe	—	prwelch.lnk	prwelch.cpp	welch.h	codebase.h
				strucpw.h	—
				date.cpp	—
				deldup.cpp	welch.h
					putext.cpp

Task: Connection of Welch and Ambula

exe-file	makefile	link file	modules	included	included
ambwelch.exe	—	ambwelch.lnk	ambwelch.cpp	welch.h	codebase.h
				strucpw.h	—
				deldup.cpp	welch.h
					putext.cpp

Task: Connection of Welch, TGL and Red Numbers

exe-file	makefile	link file	modules	included	included
tglwelch.exe	tglwelch.wmk	tglwelch.lnk	tglwelch.cpp	welch.h	codebase.h
				strucpw.h	—
			putext.cpp	welch.h	—
			date.cpp	—	—
			deldup.cpp	welch.h	codebase.h
				putext.cpp	welch.h
			index.cpp	welch.h	codebase.h

Task: Creation of an empty file AMBULA.DBF (unused in the end)

exe-file	makefile	link file	modules	included	included
ambula.exe	—	ambula.lnk	ambula.cpp	welch.h	codebase.h
				ambula.h	—

Table 12: Summary of building the programs for data manipulations