Data preparation

This file gives a report on how the original data were brought into shape for the analysis for the paper Grosche, B., H. Katayama, M. Hoshi, K. N. Apsalikov, T. Belikhina, Y. Noso and N. Takeichi (2017). "Thyroid diseases in populations residing near the Semipalatinsk Nuclear Test Site, Kazakhstan: Results from an 11 years series of medical examinations." SM J Publ Health Epidemiol 3(1).

Q: indicates questions B Grosche sent to partners

A: indicates their answers

Original file: 20150910_data(01)

Modified file: 20150910_data(02)

Modifications:

column "Exposed location"

- Transcription from Cyrillic, partly translation (e.g. for names of months), using Google translater

Modified file: 20150910_data(03)

- Introduction of the column "exposed (BGr)" after "Exposed location": If people lived in an unexposed area and moved into the exposed area only after the relevant test, they were classified as unexposed; if people were born after the relevant test (even only one day later), they were classified as unexposed;
- Codes used are:
  - x – exposed
  - n – not exposed (either living not in an exposed settlement or born after the relevant test)
  - q – questionable
  - s – Semipalatinsk (because of debate to what extent the city was affected by the 1949 test)
- exposed places of residence are:
  - see table below
- Table of settlements in the data set, broken down by exposed, not exposed, and unclear exposure situation (not including Semipalatinsk)

EXPOSED	**NOT EXPOSED**	**UNCLEAR**
Bolshaya Vladimirovka	China	B. Kamenka
Cheremuski	Mongolia	Degelen
Dolon	Russia	M.Vladimirovka
Kainar	Ukraine	Pavlodar
Karaul	Uzbekistan	(empty field)
Kunduzdy		
Mostik	Abay	Keruzova
Sarzhal	Abraly	Kokbai
Znamenka	Akatal	Paladolynysky

	Aksuat	Semey
	Aktogay	
	Algabas	
	Alma Ata region	
	Altoz	
	Aul	
	Ayaguz	
	Baladzhal	
	Baladzhal	
	Baskhkul	
	Begen	
	Belokamenka	
	Beryozovka	
	Beskaragai	
	Blagoveshchensk	
	Bolshevik	
	Buras	
	Buras Ramadan	
	Chzunbulak	
	Galago	
	Georgievka	
	Gluhovka	
	Grachi	
	Kanbay	
	Kara Marsha	
	Kokpekty	
	Koyanbay	
	Krivinka	
	Krivinka	
	Kurchum	
	Kurchum	
	Kyzyl Tu	
	Kzyl-Tu	
	Morkovka	
	Novoshulba	
	Novoshulba	
	Orenburg	
	Petropavlovka	
	Petropavlovsk	
	Rudnik Baladzhal	
	Samiyarka	
	Samiyarka	
	Scherbakty	
	Semenovka	
	Semey-Birlik	
	Semiyarka	
	Semyonovka	
	Shelekovo	
	Shemonaiha	
	Socialistik	
	Sosnovka	

	sovkhoz Azbulak	
	sovkhoz Chagan	
	sovkhoz Ondurus	
	sovkhoz Socialistik	
	Taylan	
	Tekeli	
	Undrus	
	Urdzhar	
	Urdzshar	
	Ust-Kamenogorsk	
	Zaisan	
	Zhaly	
	Zhanazol	
	Zhangiz-Tobe	
	Zharma	

checked exposure:

- Kokpekty (not exposed, see Bauer et al., 2005),
- Scherbakty (not exposed, outside the trace of the 1949 test, which passed over Rubtsovsk, Russia),
- Bolshaya Vladimirovka (exposed, see Land et al., 2008),
- **But what about M. Vladimirovka (sysid 51466)?**

Q1: Place of residence

Meaning of ※1, …, ※5 ?

A1:

※1：1976 - 1983 (Semipalatinks) → (Kachirusky area, Baudarskaya) – *corrected in data set 06.10.15*
※2：1947 – 1952 (Beskaragaysky) → 1963〜1974 (Semipalatinsk) → (Jeskent village, Baradoika region) – *corrected in data set 06.10.15*
※3：1957 – 1972 (Beramenka village, Beskaragaysky) → Semipalatinsk – *data set said "Belokamenka" instead of "Beramenka"; I couldn't find Beramenka, so I left Belokamenka*
※4：1948 – 1970 (Jana-Semeiky) → 1970 – 1996 (Jarminisky) → 1996 – (Kurchatov)
※5：1958 – 1978(Karaur) → Semipalatinsk

Meaning of "?" → A: Unknown

Meaning of empty cells (here named "empty") → A: Unknown (There are no medical records, but only the result of chemical examination)

Meaning of "Dolon ?" A: ---------------→ "Dolon"

Place of residence	Frequency	Exposed
?	1	?
※1	1	?
※2	1	?
※3	1	?
※4	1	?
※5	1	?
Beskaragai	9	NO
Beskaragysky	2	NO
Dogolan	1	NO
Dolon	333	YES
Dolon？	3	?
empty	3	?
Kainar	181	YES
Karaul	43	YES
Keruzova	1	NO
Kokbai	1	NO
Kokpekty	224	NO
Paladolynysky	1	NO
Sarzhal	309	YES
Semipalatinsk	89	Probably
Sherbakty	59	NO
Socialistic	20	NO
Sosnovka	2	NO
Sum	1287	

- A: Beskaragai and Beskaragaysky are the same region.

Q2: Meaning of variable "Exposed"

1- exposed?
2- Not exposed?

Empty -?? → no information

A: This value was corrected from the questionnaire at the examination. There are no detail information about this value, and "exposed" people might have a certification from the Institute.

Q3:

Is "Degelen" (originally Дегелен) (sysid 63026, 64157, 112063) a typo for Dogalan ?

To my knowledge, Degelen is the mountain area on the test site, while Dogolan is a settlement close to Karaul and probably exposed

A: I do not have any answer about this. – *I sent a respective question to Tatjana in the Institute; 06.10.15*

Q4:

What about those individuals without a sysid?

A: Sysid was the key of the Institute's database. Therefore, no sysid means no entry in the Institute's database.

Q5:

I remember there was a way to identify people who were examined more than once. How could that be done?

A: If you sort the data by sysid, you can see the same sysids. Those people were examined more than once. I think that some other people who do not have sysid had the examination more than once, but it is very difficult to identify those people because there is no key to identify them.

Q6:

If "race" is empty – no information?

A: Yes, there is no information. The race was corrected from 2004 at the time of examination. Therefore, the race before 2004 derived from the Institute's database.

Meaning of "race2" ?

A: They are discrepancies between the data from the medical record sheets and the data from the database. At that cases, the race2 is the data from the database.

Modified file: 20150910_data(04)

According to the answers given above

Modified file: 20150910_data(04)_spss

Selected variables for descriptive analysis:

- Year
- Ex-no
- Race
- Sysid
- Death_dt
- Residence; codes

Semey	-2
Unclear (exc. Semey)	-1
Not exposed	0
Bolshaya Vladimirovka	1
Cheremuski	2
Dolon	3
Kainar	4
Karaul	5
Kunduzdy	6
Mostik	7
Sarzhal	8
Znamenka	9

- Birth-date
- Age
- Sex
    - 1 – male
    - 2 female
- Height
- Weight
- Exposed
- Exposed (BGr)
- Medicine
    - 1 = yes
    - 2 = no
    - Empty = n.a.
- FT3
- FT4
- TSH
- Thyroidism
    - -3 = ((T3 >= 0.7 AND T3 < 2.1) AND (T4 >= 4 AND T4 < 12)) AND TSH > 3.7
    - -2 = T3 < 0.7 AND T4 < 4.0
    - -1 = T3 < 0.7 OR T4 < 4.0
    - 0 = (T3 >= 0.7 AND T3 <= 2.1) AND (T4 >= 4 AND T4 <= 12)
    - 1 = T3 > 2.1 OR T4 > 12
    - 2 = T3 > 2.1 AND T4 > 12
        - Values
        - -3 = sub-clinical hypothyroidism
        - -2 = hypothyroidism
        - -1 = hypothyroidism

    - 0 = normal
    - 1 = hyperthyroidism
    - 2 = hyperthyroidism
- Function (converted text from “Function of thyroid” to numbers
    - Values
    - -3 = sub-clinical hypothyroidism
    - -2 = hypothyroidism
    - -1 = slight hypothyroidism
    - 0 = normal
    - 1 = slight hyperthyroidism
    - 2 = hyperthyroidism

Question 1

According to the information on the normal ranges of T3 and T4, the above mentioned categorization would be possible. But when looking at the data, there is some conflicting information regarding T3 and T4; e.g. sysid 56716: while T3 is 2.2 and indicative for hyperthyroidism, T4 is 0.9 and indicative for hypothyroidism.

How to deal with such cases?

Should I only take the written information on “Function of thyroid” into account?

If yes, what is meant by “hyporthyroidism” – hypo- or hyper- ?

Answer from Hiro and Dr. Noso

"hyporthyroidism" was "hypothyroidism". Please collect all of them. *done*

I got the answer from Dr. Noso. Here is his explanation:

Most hormone produced from a thyroid is T4. Most of T3 is re-produced from T4 at a liver and other organs. The effect as hormone of T4 is weak, but T4 has a strong power to effect to a cell. Therefore, FT4 is used to check the ability to produce a hormone of thyroid, and FT3 is used to check the effectiveness against a whole body by thyroid hormone.

[standard]
FT4 (Free Thyroxine) 0.9 ~ 1.9 ng/dl
FT3 (Free Tri-thyronine) 2.5 ~ 4.5 pg/dl

> T3 is 2.2 and indicative for hyperthyroidism, T4 is 0.9 and indicative for hypothyroidism

This is normal.

1) use the collect standard rate.
2) There may be a people who are under treatment. Therefore, the value of T3 and T4 may exceed the standard value. If the value does not exceed so much, please set those cases as normal.

----- Here is my question to him --------
Would you please teach me how much the measured value exceeds the standard value to decide abnormal?

------ his answer --------------------------

If the measured value is more than twice of the standard rate, use "abnormal". In general, we check the value of T3, T4 and TSH, then decide whether hyper or hypo.

Data corrected accordingly on 8 Jan 2016 in file data_20150910(04)_spss only!

- Cyto (classification of "cytology"; in case of more than one diagnosis the following ranking of importance was used: thyroiditis, nodule, carcinoma. Cases with diagnoses like "possibly" were handled as if diagnosis was confirmed)
  - 0 = empty, no malignancy,
  - 1 = colloid nodule (adenomatous nodule, colloid nodular goiter)
  - 2 = follicular adenoma
  - 3 = follicular cancer
  - 4 = papillary cancer
  - 5 = benign tumor
  - 6 = struma
  - 7 = thyroiditis
  - 8 = class II / III (carcinoma)
  - 9 = metastasis
  - -9 = blood only, lymphocyte only, no cells
- Race recoded to ethnic
  - -1 = empty, -, ?
  - 1 = Kazakh
  - 2 = Russian
  - 3 = Bashkir
  - 4 = Bulgarian
  - 5 = Byelorussian
  - 6 = Georgian
  - 7 = German
  - 8 = Tartar
  - 9 = Ukrainian

Page "Total_new_spss" contains values from logical operations in "Total_new".

The following columns were deleted

- death_dt (not relevant for the analysis)
- Birth_date (age at time of examination is sufficient)
- Exposed (exposed (BGr) is based upon individual situation; renamed to "exp2")
  - Exp2 recoded as follows
    - -2 = Semey
    - -1 = questionable
    - 0 = no
    - 1 = yes
- Thyroidism (initially thought for categorizing the hormone status, but not used because of explanations by Dr. Noso)
- Function of thyroid (categorized as "Function")

Data modification

Year 2000 examinations numbered as "K-###" reformatted to ###

All empty cells filled with "-99" (missing values)

Introduction of new variable CYTO2 for multiple diagnoses:

- 0 = no diagnosis
- 1 = thyroiditis (T)
- 2 = nodule (N)
- 3 = T+N
- 4 = carcinoma (C)
- 5 = T+C
- 6 = N+C
- 7 = T+N+C

Saved as EXCEL 5.0 file (containing only the sheet “Total_new_spss”)

Corrections on cytology related codings were made in the files

- Data_20150910(04)_spss_input.xls
- Total_new.sav (SPSS system file)

SPSS

Nodoubles:

Of the overall 1287 individual examinations in the data set, 711 had a sysisd and 576 did not.

If one individual was examined more than once and this was identified by sysid, only the most recent information was kept. Thus, 557 different individuals are included in the study population (21.7% excluded).

For those 576 individual examinations with no sysid, plausibility was checked manually based upon date of birth, gender, ethnicity. Basis for this was the file data_20150910(04).xlsx. 511 individual examinations are left in the data set (11.3% excluded)

Excluded were:

Year / Ex-no	Same as
1999 105	2001 129
1999 106	2001 148
1999 109	2005 28
1999 122	2001 158
1999 152	2001 226
1999 168	2001 240
1999 171	2001 239
1999 55	2005 77 (needs to be discussed)
1999 173	2003 33
1999 98	2000 18
2000 22	2002 16 (needs to be discussed)
2001 123	2005 28
2001 130	2003 107
2001 196	2005 41 (2003 112 seems to be someone else: different por)
2001 39	2008 56 (2005 15 seems to be someone else: different por)

2001 70	2005 72
2002 18	2007 87 (could be different persons; needs to be discussed)
2002 6	2003 108
2003 1	2004 3
2003 88	2005 19 (could be different persons; needs to be discussed)
2004 32	2005 43
2006 11	2007 3
2006 44	2008 113
2006 77	2008 44
2006 84	2008 101
2007 15	2008 15 (obviously a mistake, both were included in analysis file; 09.03.2017)
2007 47	2009 8
2007 61	2008 44

I excluded all 2000 K-### cases, because there was no possibility to check for possible doublicates.

Not excluded were:

1999 8 and 1999 19: same date of birth and sex, but different height and weight

2001 69 and 2003 8: dto.

2005 49 and 2006 50: dto., and different place of residence

2001 110 and 2005 75: dto., and different residential history

2005 21 and 2005 80 and 2006 109 dto., different place of residence

2003 113 and 1999 90 same date of birth as 2005 72: dto., diff. place of residence

1999 4 and 2004 4: dto., diff. place of res.

2002 27 and 2003 86: dto., different height and weight

1999 37 and 2005 78: dto., different height and weight

1999 7, 2005 67, 1999 38: dto., different height and weight, different place of res.

2055 44, 2005 63: diff. por (place of residence)

1999 49 and 2001 59: dto., diff. place of at time of exposure (needs to be discussed)

1999 48 and 1999 50: dto., very similar, but unlikely that the team examined the same person twice during the tour and at the same date

1999 12 and 1999 13: see 1999 48 and 1999 50

1999 31, 2005 52, 2006 65 and 2006 72: all same date of birth and sex, but different por

2004 117 and 2007 31: different ethnicity

2001 118, 2005 65 and 2005 82: unclear situation, 2001 118 could be 2005 65 or 82; needs to be discussed

1999 141 and 2007 86: look alike, but size and weight different

2001 38 and 2003 92: look alike, but size and weight different

2001 31 and 2003 96: look alike, but size and weight different (needs to be discussed)

New SPSS file is nodoubles(complete).sav

It contains data for 1133 individuals, 557 with and 511 without a sysid

For easier case retrieval a new identifier is introduced: ID = (year*1000)+exno

Analyses

- Frequencies all
- thyroid cancer: 42 cases, 13 MV (system), 1120 valid information
- Hypothyroidism: 64 cases among 1133 individuals

Risk analyses hypothyroidism

Select cases if age > 34 (agegrp > 3)

Freq function2 agegrp sex expsett

Crosstabs

- Function2 * agegrp sex expsett
- Function2 * expsett * agegrp * sex
- Tests: chi² and Mantel-Haenszel
- If expsett = exposed: function2*exposed*agegrp*sex

Risk analyses thyroid carcinoma

Select cases if age > 34 (agegrp > 3)

Freq thycan agegrp sex expsett

Crosstabs

- thycan * agegrp sex expsett
- thycan * expsett * agegrp * sex
- Tests: chi² and Mantel-Haenszel
- If expsett = exposed: thycan*exposed*agegrp*sex

Finally, only exposed settlements were considered (see different health status between exposed and other settlements)

Select if agegrp > 3 and expsett = 1

Crosstabs

- Function2 thycan * exp2 agegrp sex
- Function2 thycan * exp2 * agegrp sex
- Function2 thycan * exp2 * agegrp * sex

Additional analysis for thyroid carcinoma

- Aged 45+ only
- Additionally stratified by ethnicity (ethnic2)

Additional analysis for hypothyroidism

If medication has an impact on the diagnosis and the risk is so much lower amongst the exposed than amongst the non-exposed – do the exposed get more medication? For this analysis the complete dataset is used (total_new.sav), i.e. including multiple examinations for some individuals.

Variable expsett had to be introduced

Recode residence (-2 = -2; -1 = -1; 0 = 0; 1 thru hi = 1); MV = -2,-1; 0 = no, 1 = yes

No information on medication is considered as no medication, thus

Recode medicine → medic2 (1 = 1, -99 and 2 = 2)

Crosstabs medic2 * exp2 expsett sex agegrp ethnic2

Crosstabs medic2 * exp2 expsett * agegrp * sex

24.10.2016

One question to Takeichi was the meaning of Class II and Class III. 8 cases have this classification in variable cytox, but in Takeichi's file data_20150910(04)_Takeichi_rev.xlsx no new information is included for these cases (based upon Total_new.sav). Thus, I go on with cyto2.

Calculation of prevalence by age and sex, based upon cyto2, values 1 (thyroiditis), 2 +3 (nodule), and 4-7 (carcinoma)

18.11.2016

New SPSS-system files nodoubles(complete)v2.sav

- After talk to Hiro in Lyon, 08.11.16: Semey considered as not affected (instead of missing/unclear); recoded -2 → 0
- Place of residence , value label for 0 now "not affected" instead of "not exposed" to better distinguish between affected settlements and exposed individuals

Changes also made in total_new.sav .

20.11.2016

Modifications in nodoubles(complete)v2.sav

- Exp2 (exposed); recoded -2 → 0 (Semey is not exposed)
- Variable expsett renamed to affsett – exposed settlement to affected settlement
    - Recoded -2 → (Semey is not exposed)
- Cytox: -9 (blood / lymph. only, no cells) recoded to 0 (no); according to Takeichi's comment
- Cytox and Cytox2: value label for 0: no → NAD (no abnormality detected)

22.11.2016

Modifications in nodoubles(complete)v2.sav

- Filter variables with x (for exposed settlement) replaced by filter variables with aff (for affected settlement)

09.03.2017

Evaluation for more than 1 examination

- Extraction of those with more than 1 examination from data_20150910(03) to data_20150910(03)-2+exams
- For those with sysid: keeping only those with more than 1 examination
- For those without sysid: keeping manually only those with more than 1 examination based upon table above (data_20150910(03)-2+exams(nosysid) is an intermediate working file)
- Introduction of variable "multexam" to define pairs, triples etc.