PERGAMON

European Journal of Cancer 36 (2000) 2380–2384

European Journal of Cancer

www.ejconline.com

Combining a name algorithm with a capture–recapture method to retrieve cases of Turkish descent from a German population-based cancer registry

O. Razum [a,*], H. Zeeb [b], K. Beck [c], H. Becher [a], H. Ziegler [c], C. Stegmaier [c]

[a] *Department of Tropical Hygiene and Public Health, Heidelberg University, Im Neuenheimer Feld 324, 69120 Heidelberg, Germany*
[b] *Department of Epidemiology and Medical Statistics, University of Bielefeld, Germany*
[c] *Population-based cancer registry of the Saarland, Saarbrücken, Germany*

Received 25 April 2000; received in revised form 31 July 2000; accepted 23 August 2000

Abstract

An increasing proportion of the 2 million Turkish residents in Germany is reaching the age in which cancer becomes a common health problem. However, data on cancer incidence and survival among Turkish residents are lacking due to incomplete reporting of nationality in German cancer registries. In the population-based cancer registry of the Saarland, retrieval by reported nationality yielded only 38% (95% confidence interval (CI): 31–45%) of the estimated number of Turkish cases in the registry; furthermore, nationality information was found to be inaccurate, and completeness dependent on the vital status of cases. A newly developed algorithm based on family names retrieved 85% (95% CI: 79–90%) of Turkish cases. Combining the two sources in a capture–recapture approach yielded 91% (95% CI: 86–94%) of the estimated total number of Turkish cases. Hence, the name-based algorithm provides a new and attractive tool for valid registry-based cancer research among Turks in Germany.

Keywords: Cancer; Registries; Transients and migrants; Bias; Germany; Turkey; Epidemiology

1. Introduction

Several hundred thousand Turkish 'guest workers' migrated to Germany in the 1960s and early 1970s, predominantly at young ages. Later, family members followed, and today, more than 2 million Turkish nationals reside in Germany [1]. An increasing proportion, as well as absolute number, of Turkish residents is now reaching the age in which chronic, non-communicable diseases like cancer become the predominant health problem. Information on cancer occurrence among Turks in Germany would be of interest for several reasons. Firstly, incidence data would allow the determination of how common particular malignancies are in this group, and which malignancies deserve special attention with regard to prevention, service provision and research [2]. Secondly, a comparison of cancer survival in the migrant and the host population would provide evidence of possible differentials in access to, and quality of, cancer screening and clinical care. Finally, migrant studies would create opportunities for aetiological research by comparing cancer risk between groups with differing lifestyles, e.g. between Turks in Germany and Germans; and between groups of similar genetic background living in different environments, e.g. between Turks in Turkey and Germany [3].

However, data on cancer incidence and survival among Turks in Germany are lacking. The main reason is that in German cancer registries, information on the nationality of reported cases is assumed to be inaccurate and incomplete. Moreover, a small proportion of Turks take up German nationality after a legally required minimum period of residency [4], so the criterion 'Turkish nationality' is no longer inclusive of all individuals of the population of interest. Hence, 'Turkish descent' would constitute a more comprehensive and also more relevant criterion. In many countries, place of birth is used as a proxy for descent [5], but this

* Corresponding author. Tel.: +49 6221 562578; fax: +49 6221 565037.
E-mail address: oliver.razum@urz.uni-heidelberg.de (O. Razum).

0959-8049/00/$ - see front matter
PII: S0959-8049(00)00333-6

information is not routinely recorded in German registries [6]. We here assess whether an algorithm based on family names provides a feasible means of retrieving cases of Turkish descent from a population-based cancer registry in Germany, independent of the nationality recorded. We then apply a simple capture–recapture method to estimate the total number of Turkish cancer cases in the registry, which in turn allows us to quantify (i) the degree of under-ascertainment of Turkish descent in the registry; and (ii) the association between the vital status of Turkish cases and a correct recording of their nationality in the registry.

2. Patients and methods

We conducted our study in the population-based cancer registry of the Saarland, the only registry in West Germany with continuous and fairly complete registration of incident cancer cases and cancer deaths since 1970. Table 1 shows population trends in the federal state of the Saarland by Turkish and German nationality. While the ageing of the Turkish population is clearly visible, current population size and age distribution indicate that only a small fraction of the 154 327 cancer cases in the registry will be Turkish.

Prerequisites for an accurate estimate of the number of Turkish cancer cases in the Saarland are that doctors will report all diagnosed cancer cases in the catchment area of the registry, independent of the national origin of the patient; and that Turkish residents who fall ill will visit a doctor for diagnosis and not re-migrate to Turkey undiagnosed. The overall completeness of cancer registration in the Saarland is high, ranging between 95.5% in 1970 and 96.3% in 1985 [7]. The DCO index (proportion of cases notified to the registry by death certificate only) is between 2.1 and 3.3% for the age group 15–60 years [7].

When reporting a new cancer case, the reporting doctor is expected to determine the nationality of the case. However, staff at the cancer registry believe that this information is frequently missing or inaccurate. When a case dies, the death certificate is copied to the cancer registry. It contains the nationality of the deceased according to the state population register which provides the 'gold standard'. Hence, completeness as well as accuracy of nationality assignment may be higher among the deceased than among the surviving cases, and the former would be over-represented in a search by nationality. If this were the case, proportional incidence ratios calculated on the basis of cases notified as Turkish in the registry would be biased towards lethal cancers, and survival times would be systematically underestimated.

Table 1
Population trends by nationality in the federal state of Saarland, Germany[a]

	German nationals			Turkish nationals		
	Total (*n*)	Men (%)	Per cent above 54 years of age	Total (*n*)	Men (%)	Per cent above 54 years of age
1970	1 089 849	47.2	n/a	1456	n/a	n/a
1980	1 017 616	46.9	26.1	6779	64.9	0.4
1990	1 011 940	47.8	29.1	10 558	57.6	2.5
1997	992 095	48.0	32.7	16 052	56.2	5.9

[a] Source: Statistisches Landesamt des Saarlandes. n/a, not available.

The name-based algorithm, described in detail elsewhere, allows the independent identification of cases of Turkish descent. In brief, Turkish family names were introduced only in 1934, have a meaning in the Turkish language [8] and can thus almost unequivocally be identified as Turkish. There are only a few 'doublets', i.e. family names like 'Bayer' that could be either Turkish or German. We compiled a list of 12 200 Turkish family names and of 2600 first names common in Turkey. The list was matched with the family names, first names and birth names (in women) of all cancer cases in the 'identified' part of the registry containing person-related data, and matching names were flagged (the other, 'non-identified' part of the registry contains epidemiological data without name identifiers). Ambiguous cases, as well as cases not flagged by the name algorithm but notified as Turkish in the registry, were re-assessed manually; criteria for the exclusion of Turkish descent were a non-Turkish first name (for doublets and cases with family names of uncertain origin), non-Muslim religion on the case record; and non-Turkish nationality on the death certificate unless the other criteria clearly indicated Turkish descent.

We then performed a capture–recapture analysis [9] where we denoted as 'capture' all cases flagged as Turkish by the name-based algorithm. The 'recapture' constituted of all cases recorded as Turkish in the 'identified' part of the registry, excluding those who were clearly not Turkish according to the exclusion criteria listed above. The proportion of the number of cases present ('flagged') both in the capture and recapture to the total number of cases present in the recapture is equal to the proportion of the number of cases in the first capture to the number of cases in the whole registry [9]. A simple transformation leads to equation (1) which we can use to estimate the total number of Turkish cases in the registry, $\hat{N}$ [10].

$$\hat{N} = \frac{n_{\text{flagged in capture}} \times n_{\text{flagged in recapture}}}{n_{\text{flagged both in capture and recapture}}} \quad (1)$$

We used a goodness-of-fit based method to obtain an estimate for the standard error ($\hat{SE}$) of $\hat{N}$ and to

construct a 95% confidence interval (CI) [10]. We then calculated binomial exact 95% CIs for the estimate of the proportion of cases identified by each source [11], assuming that the total number of cases in the registry, N, is known.

We quantified the association between the vital status of Turkish cases and a correct recording of their nationality in the registry by comparing the ratio of deaths to incident cases (i) among cases identified by the name algorithm; and (ii) among cases reported as Turkish in the 'identified' and/or the 'non-identified' part of the registry. We used a Chi-squared test to assess for difference in proportions. All calculations were done with Stata 6.0 statistical software [12]. The study was cleared by the state representative for the protection of data privacy.

3. Results

3.1. Identifying Turkish cases in the registry using the name algorithm

The cancer registry of the Saarland contained 154 327 cases on 31 December 1998. When matching the cases in the 'identified' part of the registry with the Turkish name lists, 1497 cases (1%) were flagged in which at least one name element was possibly Turkish. Only 1% (1/81) of cases with two Turkish name elements but 99.5% (830/834) cases flagged as 'doublets', i.e. with family names that could be either German or Turkish, were excluded as non-Turkish using the algorithm (see Table 2). In 582 cases (0.38% of all registered cases), religion and/or nationality had to be updated from other records, or the first and family name had to be inspected, before a decision on descent could be reached. Ultimately, 167 cases were identified as Turkish, 4 of these were excluded because they had been diagnosed in 1968, 2 years before the registry became fully functional. The final sample comprised 163 cancer cases of Turkish descent, equal to 0.1% of all cases in the registry; the sample included 3 cases with ICD 173 (malignant neoplasm of the skin, excluding melanoma) and 17 cases with ICD 231-238 (carcinoma *in situ* (CIS) and neoplasm of uncertain behaviour). Among the 163 cases were 98 men (60%); 146 (90%) of the cases were aged between 15 and 59 years. The DCO index was 3.1%.

3.2. Quantifying underascertainment of nationality in the registry

In the 'identified' part of the registry, 77 cases were reported as 'Turkish', 5 of these (6%) were definitely not Turkish on manual examination of their records or death certificate. Of the 72 remaining cases, 61 (85%) had also been flagged by the name algorithm. Using the capture recapture method (equation [1]), the total number of Turkish cases in the registry can be estimated as $163 \times 72/61 = 192$ (95% CI: 178–207). The name algorithm thus identified 85% (95% CI: 79–90%) of Turkish cases in the 'identified' part of the registry; it missed 29 cases (15% of the estimated total; 95% CI: 10–21%) whereas the reporting by nationality missed 120 cases (62.5% of the estimated total; 95% CI: 55–69%) and identified only 37.5% (95% CI: 31–45%). Combining the two sources yielded 91% (95% CI: 86–94%) of the estimated total number of Turkish cases.

3.3. Quantifying misclassification of nationality among deceased cases in the registry

49 deceased cases were flagged by the name algorithm. In 45 of them, manual inspection of the death certificate confirmed Turkish nationality. In 15 of these 45 cases (33%) the Turkish nationality had not been updated in the respective cancer registry records. Instead, these cases had been misclassified as German (9 cases), other European (3 cases), non-European (2 cases) or unknown (1 case) nationality, probably by the reporting doctors. The 4 remaining flagged deaths (8%) were of

Table 2
Identification of Turkish cases in the cancer registry using a name-based algorithm[a]

Category	Flagged in automatic search	Cases accepted as Turkish	Exclusion criteria and mode of decision
Most likely Turkish (two Turkish name elements)	81	80	Reported nationality neither Turkish nor German (algorithm)
Family name 'doublete'	834	4	First name not matching with list of Turkish first names (algorithm)
Possibly Turkish (one Turkish name element)	582	83	Non-Muslim religion; reported nationality neither Turkish nor German (algorithm)
			Non-Turkish family/birth name (manual)
Total	1497	167	

[a] Total number of cases in the registry: 154 327. Manual update of information and/or manual decision required in 582 cases = 0.38% of total cases.

German nationality according to the death certificate. Their Turkish names and Muslim religion indicated that they were of Turkish descent and had taken up German nationality.

3.4. Quantifying the association between reported nationality and disease outcome

Of the 163 Turkish cases identified by the name algorithm, 49 (30%) had died. Among a total of 111 cases reported as Turkish in the 'identified' and/or the 'non-identified' part of the registry, 53 (48%) had died ($P = 0.003$). This comparison indicates that a search for cases based only on the nationality information available in the registry would lead to a ratio of deaths to incident cases inflated by a factor of 1.6 (1.5 if 7 deaths reported as Turkish, but with a different nationality on their death certificate are excluded).

4. Discussion

Our study is the first to quantify the degree to which assignment of Turkish nationality in the population-based cancer registry of the Saarland is inaccurate and incomplete. The high level observed is partly due to incorrect or incomplete information from reporting doctors. In addition, the correct nationality had been updated in the registry in less than 70% of Turkish cancer deaths, even though a copy of the death certificate containing the definite nationality information had been filed. Completeness and accuracy of nationality assignment in the registry was also found to depend on the vital status of the cases. When cases registered as Turkish are retrieved, the ratio of deaths to incident cases will be overestimated by 50–60%. The relative importance of cancers with a high case fatality rate will thus be overstated and a comparison of survival time by national origin will be biased. Therefore, an independent method to identify Turkish cases is required, e.g. the name algorithm presented here.

Name algorithms have been successfully used to identify individuals of Mexican–American, Chinese and South Asian origin in North American registries [13–15]. Our study is the first to demonstrate that the use of a Turkish name algorithm is feasible in a German cancer registry. It was shown to more than double the fraction of Turkish cases that could be retrieved. Through the use of the algorithm, a manual assessment of case records and death certificates was required in only a small, easily manageable proportion of all registered cases.

Would our study approach allow a valid estimate of cancer incidence rates among Turks in the Saarland? One prerequisite is that reporting of incident cases in the catchment area of the registry is reasonably complete and independent of national origin. This appears to be the case: the DCO index in the Turkish sub-sample identified by the name algorithm compares well with that of similar age groups in the registry. While the DCO index as a measure of incompleteness of cancer registration has well-known limitations [16] it should allow a fair comparison when age groups and cancer sites are broadly similar. Next, retrieval of Turkish cases from the registry needs to be sufficiently complete. The use of the name algorithm alone will not suffice: a list of family names can never be exhaustive so it will not be possible to identify *all* Turkish cases in this way. Using reported nationality as a second independent source of information in a capture–recapture approach increases the level of completeness, in our example from 85% to 91% (174 out of the estimated 192 cases were present in either or both sources). A prerequisite for a correct estimate is that two assumptions underlying the capture–recapture method are met [9]: firstly, that being flagged in the first capture does not influence the chance of being retrieved in the re-capture; and secondly, that each Turkish case in the registry has an equal chance of being captured through the name algorithm and the recorded nationality, respectively.

The first assumption, independence of retrieval approaches, is not violated as name algorithm and nationality reporting in the cancer registry are indeed mutually independent. The second assumption, equal probability of retrieval, has been violated to some degree because surviving Turkish cases had a lower probability of being present in the re-capture than Turkish cases who had died. This leads to a small underestimate of the total number of Turkish cancer cases in the registry, in particular of the cases still alive, in the capture–recapture analysis. The (relatively few) Turkish cases who obtained a German passport would also be missing in the recapture and only be flagged by the name algorithm. The name algorithm in its current stage of development does not flag the very small number of Turkish cases with a double (hyphenated) name and those whose name had been misspelled. These limitations would lead to a small additional underestimate of the total number of cases, independent of their vital status. A future refinement of the algorithm could incorporate a phonetic transcription of Turkish names so that spelling errors would not preclude identification. For the particular purpose of differentiating between German and Turkish descent, religion can be a useful additional criterion in the algorithm.

In conclusion, the use of a name-based algorithm to identify Turkish cases in a cancer registry considerably increases case yield when reporting of Turkish nationality is incomplete. More importantly, identification of cases will be largely independent of vital status and of current nationality (which can change). A data set based on the name algorithm, while not comprising *all*

Turkish cases in the registry, will constitute a *de facto* random sample of Turkish cases; hence, this approach provides a new and unbiased way of estimating proportional incidence ratios and comparing survival time by national descent. If the same algorithm were to be applied to the population registry covering the catchment area of the registry, a population denominator could be constructed and incidence rates could be estimated. Finally, as the name algorithm constitutes an independent 'capture', it allows the estimation of the total number of Turkish cases in the registry. The combination of name algorithm and capture–recapture method described here is relevant for the population-based cancer registries of all the German federal states and for the many other registries of health-related events where information on nationality or place of birth is incomplete or not routinely reported [6].

The need for a manual review of selected cases could be further reduced, and the precision of the capture–recapture estimate of the total case number increased, by motivating reporting doctors to correctly ascertain the nationality and also the religion of cases; and by immediately updating nationality from the death certificate in all deceased cases in the registry. Ultimately, however, place of birth and date of immigration of cases, and ideally also of both parents, should be reported. This would allow migrant studies to be conducted on cancer causes, as well as an uncomplicated and unbiased analysis of cancer incidence and survival among migrants and their offspring, from German cancer registries. At present, in the absence of this information, the name algorithm in combination with a capture–recapture approach provides an attractive alternative for valid registry-based cancer research among Turkish migrants in Germany.

Acknowledgements

Data retrieval and collation of the data set of Turkish cancer cases were supported by a grant from the German Federal Ministry of Health, Kap. 1502 Titel 652 31, 1999.

References

1. Statistisches Bundesamt. *Statistisches Jahrbuch für die Bundesrepublik Deutschland*. Wiesbaden, Statistisches Bundesamt, 1998.
2. Bhopal RS, Rankin J. Cancer in minority ethnic populations: priorities from epidemiological data. *Br J Cancer* 1996, **29**(Suppl.), S22–S32.
3. Parkin DM. Studies of cancer in migrant populations. *IARC Sci Publ* 1993, Issue 123, 1–10.
4. Beauftragte der Bundesregierung für Ausländerfragen. Daten und Fakten zur Ausländersituation. Bonn, Mitteilungen der Beauftragten der Bundesregierung für Ausländerfragen, 1999.
5. Chaturvedi N, McKeigue PM. Methods for epidemiological surveys of ethnic minority groups. *J Epidemiol Commun Health* 1994, **48**, 107–111.
6. Razum O, Zeeb H. Epidemiologische Studien unter ausländischen Staatsbürgern in Deutschland: Notwendigkeit und Beschränkungen. *Gesundheitswesen* 1998, **60**, 283–286.
7. Brenner H, Stegmaier C, Ziegler H. Estimating completeness of cancer registration in Saarland/Germany with capture–recapture methods. *Eur J Cancer* 1994, **30A**, 1659–1663.
8. Jastrow O. Die Familiennamen der Türkischen Republik. Bildungsweise und Bedeutung. In Schützeichel R, Zender M, eds. *Erlanger Familiennamen-Kolloquium*. Neustadt, Aisch, 1985, 101–109.
9. International Working Group for Disease Monitoring and Forecasting. Capture–recapture and multiple-record systems estimation I: history and theoretical development. *Am J Epidemiol* 1995, **142**, 1047–1058.
10. Regal RR, Hook EB. Goodness-of-fit based confidence intervals for estimates of the size of a closed population. *Stat Med* 1984, **3**, 287–291.
11. Bernillon P, Lievre L, Pillonel J, *et al*. Record-linkage between two anonymous databases for a capture–recapture estimation of underreporting of AIDS cases: France 1990–1993. *Int J Epidemiol* 2000, **29**, 168–174.
12. Stata Statistical Software: Release 6.0. College Station, StataCorp. 1999.
13. Hazuda HP, Comeaux PJ, Stern MP, Haffner SM, Eifler CW, Rosenthal M. A comparison of three indicators for identifying Mexican Americans in epidemiologic research. *Am J Epidemiol* 1986, **123**, 96–112.
14. Huey-Huey Hage B, Oliver RG, Powles JW, Wahlqvist ML. Telephone directory listings of presumptive Chinese surnames: an appropriate sampling frame for a dispersed population with characteristic surnames. *Epidemiology* 1990, **1**, 405–408.
15. Sheth T, Nargundkar M, Chagani K, Anand S, Nair C, Yusuf S. Classifying ethnicity utilizing the Canadian Mortality Data Base. *Ethn Health* 1997, **2**, 287–295.
16. Brenner H. Limitations of the death certificate only index as a measure of incompleteness of cancer registration. *Br J Cancer* 1995, **72**, 506–510.