British surname origins, population structure and health outcomes – an observational study of English hospital records

Jakob Petersen (✉ jakob.ptsrn@outlook.com)

University College London <https://orcid.org/0000-0002-6659-7028>

Jens Kandt

University College London

Paul Longley

University College London

Research

Keywords: Biosocial research, health, geography of surnames, hospital admissions, isonymy, surname classification

Posted Date: May 21st, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-540484/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Abstract

Background

Population and social structure act as important confounders on pathways linking genotypes to health outcomes. This study examines whether the historical, geographical origins of British surnames – as markers of population structure - are associated with differential health outcomes today.

Methods

We coded the hospital admissions of more than 30 million patients in England between 1999 and 2013 to their surname origin and divided their diagnoses into 125 major disease categories. A base population dataset was constructed with patients' first admission of any kind. Age- and sex-standardised odds ratios were calculated with logistic regression using patients with ubiquitous English surnames such as "Smith" as reference. Using a data mining approach, we scanned the results for "signals", where a branch of related surname origins all had significantly higher or lower risk than the reference group. We subsequently studied the age- and sex-standardised incidence for each signal across the density of the surname origin (quintiles) as well as quintiles of area deprivation.

Results

We identified a signal with Scottish surnames (alcohol-related disorders) and three with different branches of English surnames (disorders of teeth and jaw, fractures, upper gastrointestinal disorders). For the three English surname groups, the risk was only different to patients with other surnames in the quintile with the highest density of that group. Differential risk remained when studied across quintiles of area deprivation.

Conclusions

The study shows that surname origins are associated with diverse health outcomes and may thus act as combined markers of population structure over and above area deprivation.

Background

It is well established that the distribution of surnames correlates with that of genetic population structure [1–3]. In Britain, surnames have been passed down from generation to generation for more than seven centuries [4]. When the geographical distributions of surnames are mapped with data from 19th century censuses, it becomes clear that many surnames can be traced back to very specific localities prior to the large-scale urbanisation and migration that characterise the people of the British Isles today [5]. Despite migration and mixing of populations - many of these surnames are still most common in the same heartlands as they were in the 19th century [6]. Genetic studies have found that the rarer the surname, the more likely the bearers of that surname are related [4]. This reflects that many more common surnames have multiple apparent origins, indicating different, unrelated, ancestors. Among the more widespread surnames with multiple origins are names taken from professions (e.g., Smith) or landscape features (e.g., Ford). Previous work has characterised the regionality of British surnames defining so-called isonymy regions based on the 1881 Census geography [6–8]. Isonymy regions are geographical regions whose populations bear a distinct constellation of surnames. One study identified direct correspondence between isonymy regions in 1881 and contemporary, genetic population structure [9](we will refer to isonymy regions as surname origins hereinafter). Surname geographies can reveal patterns of migration and social mobility [10]. The 1881 regional analysis does not, however, accommodate the apparent multiple origins of surnames imported to Great Britain prior to the 1881 Census – with many names borne by migrants from Ireland being prominent examples.

In theory, bearers with the same surname origin are more likely to be related genetically and culturally. How related they are and whether surname origins could be a useful marker of co-ancestry in health studies is an ongoing research concern [4], but evidence from multiple countries suggests that present day surname bearers resident in high density heartlands are more likely to be related than bearers scattered across lower density regions, where the chance of admixture is greater.

Population structure also acts as a strong confounder which continues to limit the validity of genome-wide association studies [11], and there is a need for methods that can systematically identify population structure and inform the sampling design of such studies.

The aim of the study is to test whether surname origins are associated with health outcomes and thus inform more detailed investigations untangling biological and social factors in health. We investigate whether British surname origins can be associated with health outcomes in a large database study of hospital admissions in England today. We divide the admissions of over 30 million patients into 125 major disease categories for the study. Even though administrative hospital data are sparse on contextual data, we make provisions for this in two ways. First, we develop a dose-response relationship between health outcomes and density of different surname groups. Second, we study confounding socio-economic factors by coding patients' residences to a national index of area deprivation.

Methods

We coded Hospital Episode Statistics (HES) recorded between April 1999 and March 2014 in England by the geographic origin of their surnames. The coding was carried out by National Health Services (NHS) Digital staff in their secure environment based on our instructions. The HES Patient Demographics Service (PDS) was used to link the maiden name or first recorded surname of a patient to 1 of 74 surname origins as identified in the 1881 Census using the procedures developed by Kandt & Longley 2018 and Kandt et al 2020 (Fig. 1). The full surname classification counts 77 distinct origins of which 74 were present in HES. Ethical approval was obtained from Bromley Research Ethics Committee (REC) (Reference: 13/LO/1355) and further approval was granted by NHS Digital (HES data licence reference is DARS-NIC-28051-Q3K7L). The anonymised HES extracts were stored and processed under secure settings.

Diagnoses were coded using the International Classification of Diseases (ICD10) system [12], which for analytical purposes was aggregated into 125 Clinical Classifications Software (CCS) categories [13]. A denominator dataset was created with the first admission (of any kind) from each of 32,860,835 patients with known residence recorded between April 1999 and March 2014. In this process, patients without surname records or non-British surnames were excluded. Only patients' first admission for each disease category was kept. In the interest of keeping specificity and avoiding dilution of any associations, the analyses were carried out in an initial "signal detection" round followed by a second round of more detailed analyses of potential signals. Age- and sex-adjusted odds ratios for each of the 125 CCS disease categories were estimated for all 74 surname origins using logistic regression. For these analyses, patients from the dominant Cluster-13, whose origin is coterminous with England as a whole, was used as a reference population. Only 94 out of the 125 CCS categories had sufficient data for analysis. "Signals" were identified from this first round where the names making up a whole branch of related surname origins had admission risks significantly above or below the reference population. The rationale for defining signals in this way was that branch-level clustering suggests a lineage effect between individuals sharing surname origin and potentially also genetic and cultural roots. Age- and sex-standardised incidence per 100,000 population were then calculated, weighted according to the 2013 European Standard Population [14] of the identified signals, and broken down by origin density quintiles and area deprivation quintiles. The origin density was constructed as the percentage of patients with a given regional origin relative to all patients resident in each local authority district, separated into quintiles. Local authority was chosen the unit of analysis because all the signal origins would have an unbroken, non-zero, distribution. A dispersion ratio for each origin aggregate was calculated as Q1 divided by Q5, so that a fully dispersed origin would have a ratio of 1 and the least dispersed a ratio approaching 0. Area deprivation quintiles were coded at neighbourhood level (2011 Middle Layer Super Output Area) [15].

Results

Associations between 125 CCS disease categories and 74 different surname origins studied and plotted during the screening round. The denominators for the analyses are shown in Table 1 and the numerators (case patient numbers) for each signal in Table 2. The case numbers varied from 6,665 teeth and jaw disorder patients with Branch-44 (Southern) surnames to 25,218 fracture patients with Branch-34 surnames (Northern) (Table 2). The Scottish origin group was more evenly dispersed with a dispersion ratio (Q1/Q5) of 0.42 compared to 0.09–0.10 for the English origin groups (Table 1).

Table 1

Patient denominator characteristics by signal set surname origins, reference surname origin (Cluster-13), and total.

	Branch-6 sumames (Scottish)	Branch-44 sumames (Southern)	Branch-34 sumames (Northern)	Branch-15 sumames (Northern)	Reference Cluster- 13 sumames (England)	Total
Sex						
Male	486,829	48,535	200,665	59,059	5,561,932	15,106,398
Female	577,531	56,770	234,030	68,495	6,526,510	17,754,437
Age (years)						
< 20	255,865	23,561	96,366	29,317	2,731,390	7,424,172
20–39	275,347	23,687	99,405	29,708	2,817,877	7,722,194
40–59	253,322	23,911	99,843	30,390	2,770,088	7,536,888
60–79	213,403	24,854	102,800	29,160	2,786,014	7,525,759
80+	66,423	9,292	36,281	8,979	983,073	2,651,822
Origin density (quintile)^a						
1. Least common	136,325	6,403	27,114	8,030	-	-
2.	186,327	10,148	33,269	11,089	-	-
3.	202,629	10,251	36,530	11,505	-	-
4.	217,375	12,907	51,626	12,203	-	-
5. Most common	321,704	65,596	286,156	84,727	-	-
Dispersion ratio (Q1/Q5)	(0.42)	(0.10)	(0.09)	(0.09)	-	-
Area deprivation (quintile)						
1. Most deprived	226,294	16,966	97,508	32,676	2,457,262	6,516,508
2.	209,545	26,267	91,076	28,622	2,394,096	6,478,882
3.	204,209	25,836	86,115	24,266	2,435,752	6,659,466
4.	208,557	19,773	86,932	22,675	2,453,281	6,725,002
5. Least deprived	215,755	16,463	73,064	19,315	2,348,051	6,480,977
Total	1,064,360	105,305	434,695	127,551	12,088,442	32,860,835
a) Origin density is specific to the signal and not shown for patient population totals.						

Table 2
Patient characteristics by signal (disease category-surname origin) and totals in number of patients.

	Alcohol-related disorders		Disorders of teeth and jaw		Fractures		Upper gastrointestinal disorders	
	Branch-6 surnames (Scottish)	All other surnames	Branch-44 surnames (Southern)	All other surnames	Branch-34 surnames (Northern)	All other surnames	Branch-15 surnames (Northern)	All other surnames
Sex								
Male	7,831	170,630	2,945	756,995	11,782	1,030,431	5,542	1,203,775
Female	3,896	95,561	3,720	958,421	13,436	1,150,123	5,886	1,282,559
Age (years)								
< 20	1,574	39,754	2,169	584,483	4,700	430,858	441	101,273
20–39	2,989	68,362	2,484	663,029	4,295	364,617	1,541	322,634
40–59	4,993	110,645	1,244	299,598	3,977	335,095	3,932	827,616
60–79	2,010	43,360	653	142,602	5,629	484,692	4,545	981,616
80+	161	4,070	115	25,704	6,617	565,292	969	253,729
Origin density (quintile)^a								
1. Least common	1,366	51,818	336	432,728	1,783	443,771	584	503,411
2.	1,968	55,109	470	374,574	2,425	450,592	834	558,909
3.	2,025	47,264	531	266,907	2,462	399,554	781	440,546
4.	2,132	46,467	622	275,937	3,533	459,611	857	391,981
5. Most common	4,236	65,533	4,706	365,270	15,015	427,026	8,372	591,487
Area deprivation (quintile)								
1. Most deprived	4,151	86,380	1,276	384,103	5,881	444,551	2,969	512,185
2.	2,871	61,412	1,946	371,057	5,421	433,988	2,665	494,342
3.	1,885	47,610	1,602	343,270	4,979	444,085	2,163	499,157
4.	1,521	38,968	1,067	327,171	4,626	434,238	2,061	508,585
5. Least deprived	1,299	31,821	774	289,815	4,311	423,692	1,570	472,065
Total	11,727	266,191	6,665	1,715,416	25,218	2,180,554	11,428	2,486,334

The full results of the initial signal detection round can be found in the Supplementary Materials Tables S1 (examples with variable labels) and S2 (all signals). The hospitalisation odds ratios for the signals are shown in Fig. 2. The age- and sex-standardised incidence for each signal disease are shown broken down by origin density and area deprivation quintile in Fig. 3.

Discussion

The main objective of this research is to examine whether the origin of patients' surnames in 1881 act as plausible markers of population structure that are associated with health outcomes today. Given that most origin groups are relatively sparsely powered, we decided to screen the results for signals where an entire branch of related origins had significantly higher or lower risk than the reference population of pan-English surname bearers.

We found four such signals. All the signals were associated with greater (dis-)advantage in the high-density regions compared to the low-density regions. For the signal with alcohol-related disorders and Branch-6 (Scottish) surnames, this was also the case for patients with any other surname origin. For the three other signals, the (dis-)advantage was only significant in the highest density quintile.

The analysis of area deprivation showed that disadvantage increased with area deprivation relative to patients with all other surname origin for alcohol-related disorders and teeth and jaw disorders. For fractures and upper gastrointestinal disorders, the relative (dis-)advantage changed little across deprivation quintiles.

The results for the three English signals suggest that it is possible to detect differential health outcomes linked to the ancestral "heartland" and that the effects persist even when considering variation across different levels of area deprivation.

The patients with Scottish surnames were more dispersed than the three English groups. The fact that we did not find a density effect could be that the heartland, in this case Scotland, was outside the study area.

From a genetic perspective, it could be hypothesised that the higher risk was caused by inbreeding depression or genetic drift. The fact that the identified signals are not headline diseases with a big impact on patients' lives or healthcare budgets, could arise because they are exactly marginal diseases under low selection pressure. Inbreeding depression is classically associated with consanguineous relationships and monogenic diseases, but nascent research suggest that it can also be associated with complex diseases and other complex traits even in population-based samples [16, 17].

From a social science perspective, an alternative explanation could be that the surname classification is likely to capture a wide range of variables associated with "nurture", which may again differ between heartlands and the "host" regions. The area deprivation index used in this study is admittedly a very reductionist measure conflating diverse conditions and experiences that can lead to poorer health outcomes. Further research should therefore include more detailed data on individual socio-economic factors.

As posited by dual inheritance theory, genetic and cultural roots are intertwined [18]. While genetic inheritance can only be vertical, cultural inheritance can both be horizontal and vertical. Cultural practices can thus be passed down vertically from forbears or quickly adopted from peers horizontally. Wealth accumulated in families can be an example of vertical, cultural transmission. Which set of factors, vertical or horizontal, genetic or cultural, are dominant will depend on the particular aetiology of the health condition in question.

A motivation for further studies would either be to contextualise surname origins as a new geography or to study the fine-scale population structure in case it has implications for the design of genetic studies [19]. The former approach should refine the regional geography used here in order to accommodate the likely effects of migration of family groups prior to the collection of censuses that have been digitally encoded.

A number of limitations should be acknowledged. HES was created for administrative and billing purposes, but the data have been validated for research [20, 21]. It cannot be ruled out that there could be sub-national coding differences especially when studying the entire range of diagnoses as in this case. The purpose of a medical classification such as CCS is to break down the analyses into meaningful categories. As with any categorisation, important variation may be lost and vary depending on the system deployed.

The base population was created from HES itself, i.e., as patients with any diagnosis by surname origin. In this way it was possible to study individual surname origin groups against the dominant group with ubiquitous English surnames. The results may be biased if the patient populations are not generalisable to the populations with the same surname origin and we were not able to validate this aspect of the analyses due to the lack of external reference population data.

To avoid distorting representations of health conditions associated with multiple admissions, we only used first admission for each condition and for each patient in the base population. With data collection spanning fifteen years, there could be a bias towards exposures in patients' younger years. We assumed that any potential biases from this source would cancel each other out although there could be residual net bias if the age of onset varied markedly between the surname origin and the reference population. In addition, HES does not contain identifiers for households, which for our analyses may mean that there could be residual clustering at household level, e.g., poisoning incidents involving multiple household members with identical surname origin.

A patient may change surname, e.g., following marriage, and end up being coded to a different surname origin than the one at birth. This would make the dataset noisier, but we reduced this by only using the maiden or first-recorded surname for each patient. In addition, local intermarriage is still common and thus it can be expected that the newly adopted surname may belong to the same or a closely related surname origin [22].

Retirement migration may mean that exposures causing some health problems are systematically attributed to retirement regions. This is a limitation, although in this study it would only apply if the retiree moves into a different area deprivation quintile since all other variables would remain the same.

We studied 94 different disease categories across 74 surname origins. Especially for the first round, this constitutes a multiple comparison problem. We decided to focus on signals where groups of surname origins differed from the reference population on the basis that "lineage" would provide a higher degree of plausibility and crucially reduce the number of simultaneous analyses. In the Supplementary Materials, we have provided the results both with and without Bonferroni corrections for 94 simultaneous comparisons.

Conclusion

The study shows that surname origins are associated with diverse health outcomes and may thus act as combined markers of biosocial population structure over and above area deprivation. If related surname origins were studied - assuming either genetic or cultural lineage effect - then again it was possible to find correlates between certain health outcomes in the areas with the highest density of the surname origin aggregate. The fact that correlates were only present in the highest density quintile suggests that ancestral heartlands vary in some aspect from the rest of the country. Hypothetically this pattern could be explained by a combination of factors related to both nature and nurture, in part depending on the nature of the health outcome itself. The results may thus inform more detailed investigations untangling biological and social factors in health.

Abbreviations

CCS - Clinical Classifications Software

HES - Hospital Episode Statistics

ICD10 - International Classification of Diseases, version 10

NHS - National Health Services

PDS - Patient Demographics Service

Q1 – quintile 1

Q5 – quintile 5

Declarations

Ethics approval and consent to participate

Ethical approval was obtained from Bromley REC (Reference: 13/LO/1355). The study was conducted in accordance with relevant guidelines and regulations. As for other non-identifiable database studies, it was not practical nor desirable to obtain consent from each patient. The study only used routinely collected, secondary data and as such involved no experimental components requiring additional protocols and approvals. The HES data licence reference is DARS-NIC-28051-Q3K7L.

Consent for Publication

Not applicable.

Availability of data and material

All the data that support the findings of this study are available from NHS Digital subject to ethnical and scientific approval of a study protocol. Researchers wishing to get access to the study data can visit this website: <https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/hospital-episode-statistics>. The authors of this study had no special access privileges others would not have.

Competing interests

None

Funding

The UK Economic and Social Research Council is acknowledged for its support for the UCL Consumer Data Research Centre (CDRC) enabling this research (Grant ES/L011840/1). The funder had no direct role in relation to the specific study.

Authors' contributions

All authors contributed to conception and design, critically revised the manuscript, gave final approval, and agreed to be accountable for all aspects of work ensuring integrity and accuracy (JK, PAL, JP). JK and PAL contributed to data acquisition. JP contributed to analysis and the first draft of the manuscript.

Acknowledgements

Not applicable.

References

1. Jobling MA. In the name of the father: surnames and genetics. *Trends Genet TIG*. 2001;17:353–7.
2. Lasker GW. A coefficient of relationship by isonymy: a method for estimating the genetic relationship between populations. *Hum Biol*. 1977;49:489–93.
3. Lasker GW. Surnames and Genetic Structure [Internet]. Cambridge: Cambridge University Press; 1985 [cited 2021 May 5]. Available from: <https://www.cambridge.org/core/books/surnames-and-genetic-structure/DBB537163D65F1D6DAEACADD83EE1C6D>
4. King TE, Jobling MA. What's in a name? Y chromosomes, surnames and the genetic genealogy revolution. *Trends Genet TIG*. 2009;25:351–60.
5. Winney B, Boumertit A, Day T, Davison D, Echeta C, Evseeva I, et al. People of the British Isles: preliminary analysis of genotypes and surnames in a UK-control population. *Eur J Hum Genet*. Nature Publishing Group; 2012;20:203–10.

6. Longley PA, Cheshire JA, Mateos P. Creating a regional geography of Britain through the spatial analysis of surnames. *Geoforum*. 2011;42:506–16.
7. Kandt J, Van Dijk J, Longley PA. Family name origins and inter-generational demographic change in Great Britain. *Am Geogr Soc*. 2020;110:1726–42.
8. Kandt J, Longley PA. Ethnicity estimation using family naming practices. *PLOS ONE*. 2018;13:e0201774.
9. Kandt J, Cheshire JA, Longley PA. Regional surnames and genetic structure in Great Britain. *Trans Inst Br Geogr* 1965. 2016;41:554–69.
10. Clark G, Cummins N. Surnames and Social Mobility in England, 1170–2012. *Hum Nat*. 2014;25:517–37.
11. Richardson K, Jones MC. Why genome-wide associations with cognitive ability measures are probably spurious. *New Ideas Psychol*. 2019;55:35–41.
12. WHO. ICD-10 Version:2016 [Internet]. 2016 [cited 2020 Jan 17]. Available from: <https://icd.who.int/browse10/2016/en>
13. Elixhauser A, Steiner C, Palmer L. Clinical Classifications Software 2015 (CCS) [Internet]. U.S. Agency for Healthcare Research and Quality; 2015 [cited 2021 Feb 15]. Available from: <https://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp>
14. Eurostat. Revision of the European Standard Population - Report of Eurostat's task force - 2013 edition [Internet]. 2013 [cited 2019 May 21]. Available from: <https://ec.europa.eu/eurostat/web/products-manuals-and-guidelines/-/KS-RA-13-028>
15. Department for Communities and Local Government. The English Indices of deprivation 2015. *Stat Release*. 2015;38.
16. Clark DW, Okada Y, Moore KHS, Mason D, Pirastu N, Gandin I, et al. Associations of autozygosity with a broad range of human phenotypes. *Nat Commun*. Nature Publishing Group; 2019;10:4957.
17. Johnson EC, Evans LM, Keller MC. Relationships between estimated autozygosity and complex traits in the UK Biobank. *PLOS Genet*. Public Library of Science; 2018;14:e1007556.
18. Newson L, Richerson P. Dual Inheritance Theory. *Int Encycl Anthropol* [Internet]. American Cancer Society; 2018 [cited 2021 Mar 2]. p. 1–5. Available from: <https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118924396.wbiea1881>
19. Leslie S, Winney B, Hellenthal G, Davison D, Boumertit A, Day T, et al. The fine-scale genetic structure of the British population. *Nature*. Nature Publishing Group; 2015;519:309–14.
20. Burns EM, Rigby E, Mamidanna R, Bottle A, Aylin P, Ziprin P, et al. Systematic review of discharge coding accuracy. *J Public Health Oxf Engl*. 2012;34:138–48.
21. Herbert A, Wijlaars L, Zylbersztejn A, Cromwell D, Hardelid P. Data Resource Profile: Hospital Episode Statistics Admitted Patient Care (HES APC). *Int J Epidemiol*. 2017;46:1093–1093i.
22. Cheshire J. Analysing surnames as geographic data. *J Anthropol Sci Riv Antropol JASS*. 2014;92:99–117.

Figures

Figure 1

Dendrogram of the British isonymy regions (surname origins). The map shows regions at the top of the hierarchy ($k=19$). Regions with identified signals are highlighted.

Figure 2

Hospital admission risk by surname origin for four disease-origin branch signals. The x-axis has the order of each surname origin in the dendrogram shown in Fig. 1 putting each “branch” next to its distal sub-branch.

Figure 3

Incidence by origin density and area deprivation quintiles.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [ljhgwbrsuppltext.docx](#)
- [ljhgwbrsupplxl.xls](#)