



Smoker, ex-smoker or non-smoker? The validity of routinely recorded smoking status in UK primary care: a cross-sectional study

Journal:	<i>BMJ Open</i>
Manuscript ID:	bmjopen-2014-004958
Article Type:	Research
Date Submitted by the Author:	30-Jan-2014
Complete List of Authors:	Marston, Louise; University College London, Primary Care and Population Health Carpenter, James; London School of Hygiene & Tropical Medicine, Medical Statistics; MRC Clinical Trials Unit, Walters, Kate; University College London, Primary Care and Population Health Morris, Richard; University College London, Primary Care and Population Health Nazareth, Irwin; University College London, Primary Care and Population Health White, Ian; Medical Research Council, Biostatistics Unit Petersen, Irene; University College London, Primary Care and Population Health
Primary Subject Heading:	General practice / Family practice
Secondary Subject Heading:	Smoking and tobacco, Research methods
Keywords:	PRIMARY CARE, STATISTICS & RESEARCH METHODS, EPIDEMIOLOGY

SCHOLARONE™
Manuscripts

1
2
3 **Title: Smoker, ex-smoker or non-smoker? The validity of routinely**
4 **recorded smoking status in UK primary care: a cross-sectional study**
5
6
7
8
9

10 **Authors**

11 Louise Marston¹

12 James R Carpenter^{2, 3}

13 Kate R Walters¹

14 Richard W Morris¹

15 Irwin Nazareth¹,

16 Ian R White⁴

17 Irene Petersen¹

18
19
20
21
22
23
24
25
26
27
28
29
30
31
32 ¹Department of Primary Care and Population Health, University College London,
33 Rowland Hill Street, London, NW3 2PF, United Kingdom

34
35
36 ²Department of Medical Statistics, London School of Hygiene and Tropical
37 Medicine, Keppel Street, London, WC1E 7HT, United Kingdom

38
39
40
41 ³ MRC Clinical Trials Unit, Kingsway, London

42
43
44 ⁴MRC Biostatistics Unit, Cambridge Institute of Public Health, University Forvie
45 Site, Robinson Way, Cambridge, CB2 0SR, United Kingdom
46
47
48

49
50 Please address correspondence to:

51 Dr Louise Marston

52
53
54
55 Department of Primary Care and Population Health
56
57
58
59
60

1
2
3 University College London
4

5 Rowland Hill Street
6

7
8 London
9

10 NW3 2PF
11

12 United Kingdom
13

14
15
16
17
18 I.marston@ucl.ac.uk
19

20 Telephone: +44 20 7794 0500 (36768)
21

22 Fax: +44 20 7794 1224
23

24
25
26 Words: 3,166
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

ABSTRACT

Objectives: To investigate how smoking status is recorded in UK primary care; to evaluate if multiple imputation (MI) of smoking status can give results consistent with health surveys.

Setting: UK primary care and a population survey conducted in the community.

Participants: We identified 354,204 patients aged 16 or over in The Health Improvement Network (THIN) primary care database registered with their general practice 2008-2009 and 15,102 individuals aged 16 or over in the Health Survey for England (HSE).

Outcome measures: Age-standardised and, age-specific proportions of smokers, ex-smokers and non-smokers in THIN and the HSE before and after multiple imputation (MI). Using information on time since quitting in the HSE, we extrapolated when ex-smokers may be considered as non-smokers in primary care.

Results: In THIN, smoking status was recorded for 84% of patients within one year of registration. Of these; 28% were smokers (21% in the HSE). After MI of missing smoking data, the proportion of smokers was 25% (missing at random) and 20% (missing not at random). With increasing age, more were identified as ex-smokers in the HSE than THIN. It appears that those who quit before the age

1
2
3 of 25-30 years were less likely to be recorded as an ex-smoker in primary care
4
5 than people who quit later.
6
7
8
9

10 **Conclusions:** Smoking status is relatively well recorded in primary care.
11
12 Misclassification of ex-smokers as non-smokers is likely to occur in those quitting
13 smoking at an early age and/ or a long time ago. Those with no smoking status
14 information are more likely to be ex or non-smokers versus smokers.
15
16
17
18
19

20 21 22 **STRENGTHS AND LIMITATIONS OF THIS STUDY** 23

- 24 • This study includes data from 'real' life primary care electronic records
- 25 • First study to compare the definition of smoking status in primary care
- 26 versus a population survey
- 27 • Study focuses on data recorded in the first year after patient registration
- 28 and may not be applicable to other times.
- 29
- 30
- 31
- 32
- 33
- 34
- 35
- 36
- 37
- 38

39 **KEYWORDS:** recording of smoking, primary care databases, Health Survey for
40 England, missing data, multiple imputation
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

INTRODUCTION

A fifth of the British adult population are smokers [1] and there is still a need for further research into smoking and smoking related diseases including coronary heart disease and stroke, respiratory diseases and cancers. Routinely collected smoking data can be used in clinical practice to identify populations at risk of smoking-related diseases, such as identifying smokers to have spirometry testing to identify those with Chronic Obstructive Pulmonary Disease (COPD), or to be invited for smoking cessation services. It is important to understand the accuracy of the data, and whether cases may be missed in those with no recorded smoking status. Electronic health records, including primary care databases, have proved to be very powerful resources for epidemiological and health research.[2-12] In order to conduct such research, it is important to understand how smoking status is recorded in primary care. There is evidence that the recording of smoking status has improved substantially in UK primary care[13, 14] and most general practices now routinely record smoking status at regular intervals as a part of the Quality Outcome Framework.[15] However, we do not know how the different and non-standardised classifications of ex, non and current smokers in primary care records compare to the standardised recording of smoking status in population surveys such as the Health Survey for England (HSE).

In addition, a proportion of patients still lack a smoking status record in their primary care records. It is unclear how to deal with these patients when

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

conducting research where smoking status is either the outcome of the research or an explanatory factor for patients' health.[3, 6, 16, 17] Methodological research has demonstrated that including only patients with complete records can substantially bias the results.[18] In recent years, efforts have been made to address missing data in primary care databases[3, 17, 19] using multiple imputation, though reporting on the comparability of the results of multiple imputation with population data has been sparse. Therefore, it is unclear whether multiple imputation accurately replicates data representing the population.[3, 6, 17, 20] Our previous work on missing data in The Health Improvement Network (THIN) primary care database showed that many health indicator measurements (for example, weight and blood pressure) recorded within the first year of patients' registration at a general practice were comparable with large external datasets before and after multiple imputation.[16] However, smoking status was not directly comparable with data from the Health Survey for England (HSE). Although the proportion of smokers was similar between THIN and the HSE *before* multiple imputation of data in THIN, the proportion of smokers was substantially higher *after* multiple imputation in THIN. On the other hand, the proportion of ex-smokers was substantially lower in THIN both *before* and *after* imputation compared to the HSE. This suggests that current smokers may be adequately identified using primary care data and most people with missing data on smoking status are likely to be either ex or non-smokers. This has clinical importance as smoking status (including ex-smoking) may be used to

1
2
3 identify those at risk of disease, for example chronic obstructive pulmonary
4
5 disease or cardiovascular disease.
6
7
8
9

10 In this study we further investigate recording of smoking status in primary care
11 and explore potential reasons for the discrepancy in the proportion of ex-smokers
12 between primary care records and the HSE. Specifically, we seek to deduce
13 when ex-smokers may not be recorded as such in primary care records based on
14 information about time since quitting in the HSE. Finally, we aim to provide a
15 practical solution for imputation of missing smoking status records in routinely
16 collected clinical data.
17
18
19
20
21
22
23
24
25
26
27
28

29 **METHODS**

30 **Study populations**

31 We used data from THIN primary care database.[21] In the United Kingdom
32 (UK) 98% of the population are registered with a National Health Service (NHS)
33 general practitioner to receive routine healthcare.[22] THIN is broadly
34 representative of all general practices in the UK in terms of age and sex of
35 patients, practice size and geographical distribution.[23] The database contains
36 information on socio demographics, symptoms, diagnoses, referrals to secondary
37 care, prescribing, results of tests and health status indicators. The data provider
38 (CSD-MR) obtained overall ethical approval from the South East MREC
39 (MREC/03/01/073) and this study was further approved by a THIN scientific
40 review committee.
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3
4
5
6 For this study we selected patients aged 16 years or over who registered with a
7
8 general practice between 1st January 2008 and 31st December 2009
9
10 (N=354,204) and we examined records from the first year after the patient
11
12 registered, hence using data up to the end of 2010. Many people have a “new
13
14 patient check” soon after registration, where information on demographics, health
15
16 indicators and disease status is collected.
17
18
19
20

21
22 We compared the distribution of smoking status with that in the HSE from 2008
23
24 for those aged 16 years or over (N=15,102). The HSE is a national annual cross
25
26 sectional interview based survey of approximately 22,000 people.[24] The
27
28 survey includes questions on socio demographics, general health and
29
30 information on smoking status. The HSE has nearly complete records of
31
32 smoking (99.3%) and we therefore used the data from patients with complete
33
34 smoking information.
35
36
37
38
39
40

41 **Definition of smoking status**

42
43 In THIN, smoking status was recorded by self-report. In many general practices
44
45 this would be on the basis of a questionnaire submitted at the time of registration,
46
47 whereas in other general practices this would be recorded in conjunction with a
48
49 clinical consultation with the general practitioner or practice nurse. Patients
50
51 would be classed as current non-smoker, or current smokers. In some instance
52
53 the non-smokers would be classified as ex-smokers but this was variably defined
54
55
56
57
58
59
60

1
2
3 from one practice to another. In the HSE, smoking status was defined on the
4 basis of a series of questions (see Appendix 1) and individuals who had ever
5 smoked (but did not smoke at the time of the interview) would be defined as ex-
6 smokers, regardless of their age at quitting and length of time since they quit.
7
8 The HSE holds information on when ex-smokers quit so that age at the time they
9 quit can be deduced, whereas this information was not consistently available in
10 THIN.
11
12
13
14
15
16
17
18
19
20
21

22 **Statistical analyses**

23
24 Initially, we examined smoking status (smoker, ex-smoker, non-smoker or
25 missing) in THIN and the HSE, overall, by age group, gender and Index of
26 Multiple Deprivation 2004 (IMD) quintile[25]. Then we used multiple imputation
27 to impute missing smoking status in THIN. Multiple imputation is a statistical
28 method which uses the data available to model the likely distribution of missing
29 data.[18] A number of imputed datasets are produced in each of which plausible
30 values are drawn from the imputation model. The method is designed to
31 correctly reflect the uncertainty surrounding the missing values. With an
32 appropriate imputation model, multiple imputation is an unbiased method of
33 accounting for missing data. It is usually performed under the missing at random
34 (MAR) assumption, but it may also be performed under specific missing not at
35 random (MNAR) assumptions. These methods have been described in greater
36 detail elsewhere.[18, 26-28]
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 After preliminary analysis,[26] we included the following variables in the multiple
4 imputation models: age in years, gender and IMD quintile,[25] health indicators:
5 smoking status (three categories, non, ex and current smoker), height, weight,
6 systolic and diastolic blood pressures and disease indicators: type II diabetes,
7 coronary heart disease (CHD) and cerebrovascular accident (CVA). Multiple
8 imputation was performed using Chained Equations using the ice command
9 using Stata 11.[29, 30] Continuous variables were imputed using multiple linear
10 regression, smoking status using multinomial regression and IMD quintile using
11 ordered logistic regression. Percentages in each smoking category were
12 obtained using Rubin's Rules.[31] In the first multiple imputation we assumed
13 that smoking data were MAR and hence allowed imputed smoking data of either
14 smokers, non-smokers or ex-smokers (using a MAR assumption; hereafter
15 referred to as MAR MI). In the second multiple imputation we assumed that all
16 smokers had been recorded (so that smoking data were MNAR) and we imputed
17 missing smoking data as either ex-smokers or non-smokers (hereafter referred to
18 as MNAR MI).

19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43 Following multiple imputation we carried out age-specific direct standardisation
44 using the HSE as the standard population and the age-specific proportion in each
45 smoking category from THIN. This was done to account for the fact that the
46 mean age in the HSE was 49 years while the mean age in THIN was 38 years in
47 the year after registration.
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 We deduced the average time after which an ex-smoker is no longer classified as
4 an ex-smoker in primary care records by combining information from the HSE on
5
6 an ex-smoker in primary care records by combining information from the HSE on
7
8 when ex-smokers quit and the age-specific distribution of ex-smokers in THIN
9
10 after imputation of non and ex-smokers. This was done by ranking the
11
12 individuals in the HSE in accordance to the length of time since they quit by 10
13
14 year age groups and then 'reclassifying' individuals who had quit the longest time
15
16 ago within each age group from ex to non until we reached the same proportion
17
18 of ex-smokers in the HSE as in THIN.
19
20
21
22
23

24 **RESULTS**

25
26 In total, 354,204 individuals were included from THIN and 15,102 from the HSE.
27
28 Individuals in THIN were, on average 11 years younger than those in the HSE
29
30 (38 years versus 49 years, respectively) (Table 1). Smoking status was recorded
31
32 for 84% in THIN within one year of initial registration. Before multiple imputation
33
34 of missing data, a greater proportion of people were recorded as smokers in
35
36 THIN than the HSE (24% versus 21% respectively), and the proportions of ex-
37
38 smokers and non-smokers differed substantially between THIN and the HSE
39
40
41 (Table 1).
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Table 1: Summary statistics for THIN in the first year of registration and the HSE 2008

Variable	THIN		HSE	
	n	%	n	%
Male	164,085	46	6,760	45
Female	190,119	54	8,342	55
Missing sex		0		0
Non-smoker	165,618	47	7,874	52
Ex-smoker	49,874	14	3,966	26
Current smoker	83,526	24	3,158	21
Missing smoking status	55,186	16	104	1
Age years mean (SD)	38	(17)	49	(19)
Missing age		0		0

Abbreviations: HSE Health Survey for England 2008; THIN The Health Improvement Network.

Our first analyses used missing as a separate category of smoking, so we refer to those with reported smoking status as “known smokers” and “known ex-smokers”. The proportion of known smokers by age group was similar in THIN and the HSE between 30 and 79 years, but this was not the case for the proportions of known ex-smokers and non-smokers (Figure 1). In the HSE, the proportion of known ex-smokers increased from 12% within the 20-29 age group to 46% in the 80-89 age group. In THIN, the proportion of known ex-smokers also increased with age although the overall proportion of known ex-smokers was smaller than in the HSE for all age groups after 20-29 years. Conversely, in the HSE, the proportion of non-smokers decreased slightly from 56% in the 20-29 age group to 48% in the 80-89 age group. Within THIN, the proportion of known non-smokers remained constant with increasing age at around 43%. The proportion of missing smoking data in THIN was relatively constant at less than

1
2
3 20% until the 70-79 years age group, but increased substantially thereafter
4
5 (Figure 1).
6
7
8
9

10 (Figure 1 here)
11
12
13

14
15 In THIN, the percentage of non-smokers was greater for women (52%) than men
16 (40%) while the percentage of smokers was smaller for women (21%) than men
17 (27%). There were similar trends in the HSE, although the percentage
18 differences between sexes were smaller (smokers: 22% of men versus 20% of
19 women).
20
21
22
23
24
25
26
27
28

29 The proportions in each smoking status category varied substantially by social
30 deprivation in both THIN and the HSE (Figure 2). In THIN, the percentage of
31 non-smokers decreased from 52% in the least deprived quintile to 40% in the
32 most deprived quintile. The percentage of ex-smokers decreased slightly with
33 increasing deprivation. In contrast, the percentage of smokers increased with
34 increasing deprivation from 16% in the least deprived quintile to 34% in the most
35 deprived quintile (Figure 2). The patterns were similar in the HSE although the
36 proportion of ex-smokers was substantially larger across all levels of deprivation
37 in the HSE compared to THIN.
38
39
40
41
42
43
44
45
46
47
48
49

50
51
52 (Figure 2 here)
53
54
55
56
57
58
59
60

Analyses imputing missing smoking status

After MAR MI of THIN, age-standardised smoking prevalences still differed somewhat between THIN and the HSE. For example, 22% were ex-smokers in THIN compared with 26% in the HSE; 25% were smokers in THIN, compared with 21% in the HSE (Table 2).

After MNAR MI of THIN (that is, regarding missing values as either ex-smokers or non-smokers), the age-standardised prevalence of smoking in THIN was similar to that in the HSE (Table 2). However, the age-specific prevalence of ex-smokers was still greater in HSE than in THIN. Age-specific analysis showed that this difference was greatest at older ages, and indeed reversed at younger ages. This suggested that individuals who had quit in the less recent past might be classified as non-smokers in THIN but as ex-smokers in HSE. (Figure 3).

Table 2: Percentages within each smoking status for THIN and the HSE 2008 after various adjustments

Category	THIN			HSE	
	Complete case	After MAR MI ^{ab}	After MNAR MI ^{ac}	Observed	Reclassifying ex-smokers ^d
	%	%	%	%	%
Non-smoker	55	53	57	53	57
Ex-smoker	17	22	23	26	22
Smoker	28	25	20	21	21

Abbreviations: HSE Health Survey for England 2008; THIN The Health Improvement Network.

^a Directly standardised using the HSE age distribution as standard.

^b Imputed assuming that missing values are smokers, non-smokers or ex-smokers

^c Imputed assuming that missing values are non-smokers or ex-smokers

^d Within each age group, reclassifying the optimum number of ex-smokers as non-smokers.

(Figure 3 here)

The median time since ex-smokers quit in the HSE varied greatly by age group (Table 3), from two years (Interquartile range (IQR): 0, 3) in the under 20s to 40 (IQR: 25, 51) years in those aged 90 or over (Table 3). Equating proportions of ex-smokers in THIN to that in the HSE data suggested the typical time-window after which patients are no longer regarded as ex-smokers in primary care, but instead regarded as non-smokers, varied with age. Thus, typically individuals who registered with a general practice when they were in their forties would no longer be recorded as an ex-smoker if they quit more than 22 years earlier (when they were between 18 and 27 years of age) (Table 3). Individuals registering in their seventies would typically no longer be recorded as ex-smokers if they quit 42 years earlier (when they were between the ages of 28 and 37 years) (Table 3). Yet, most individuals who quit after the age of 30 would still be captured as ex-smokers when they later registered with a new general practice. Using these age-specific extrapolations to reclassify ex-smokers as non-smokers in the HSE according to when they quit, we can see that the age-specific distributions of ex-smokers in THIN and the reclassified HSE are similar (Figure 3).

Table 3: Age specific centiles of time since quitting smoking in the HSE 2008

Age group	Median time since quitting (years)	Extrapolated number of years since quitting	Extrapolated age when they quit
<20	2	*	*
20-29	3	*	*
30-39	5	14	16 - 25
40-49	10	22	18 - 27
50-59	20	30	20 - 29
60-69	24	35	25 - 34
70-79	30	42	28 - 37
80-89	32	40	40 - 49
90+	40	46	44+

*Not possible to assign an optimal value for reclassification to these age groups
Abbreviations: HSE Health Survey for England 2008

DISCUSSION

The proportion of newly registered patients in THIN between 2008 and 2009 with a record of being a smoker was slightly higher than the HSE in 2008. However, the proportion of individuals recorded as ex-smokers and non-smokers differed substantially between THIN and the HSE. Overall, a larger proportion of individuals were recorded as ex-smokers in the HSE than in THIN and this increased with age. Likewise, the proportion of ex-smokers was substantially larger across all levels of deprivation in the HSE compared to THIN.

Under MAR MI there was a greater percentage of smokers (25%) and a smaller percentage of ex-smokers (22%) in THIN compared with the HSE (smokers 21%, ex-smokers 26%). However, under MNAR MI (assuming all missing data were either ex-smokers or non-smokers) slightly increased the proportion of non-smokers (57%) in THIN compared to the HSE (53%), whereas the proportion of ex-smokers (23%) was slightly lower in THIN. Moreover, the latter imputation

1
2
3 resulted in a relatively larger percentage of ex-smokers in THIN in those aged
4 under 30 years compared with the HSE. This may be because the imputation
5 model was unable to distinguish between ex and non-smokers in those age
6 groups as both are unlikely to have developed typical later onset diseases which
7 are key predictors of smoking status in the imputation model.
8
9
10
11
12
13

14
15
16
17 There may be several reasons for the discrepancy in the distribution of the
18 smoking categories between THIN and the HSE. In the HSE, the definition of an
19 ex-smoker was highly sensitive and clearly defined.[24] Thus respondents were
20 categorised as ex-smokers even if they were a trivial smoker, smoked for a short
21 period of time and/ or quit many decades ago. Also, the HSE used computer
22 aided personal interviewing; where questions were read to the respondent in a
23 standardised way from the screen and a detailed sequence of questions were
24 asked to ascertain current smoking status. In primary care, while smoking status
25 is systematically recorded in medical records, there is no detailed protocol for
26 recording smoking status and the ascertainment is thus likely to vary by how the
27 information was obtained. Many practices use self-report questionnaires at
28 registration including smoking status. Smoking status is then updated by health
29 professionals (general practitioners and/ or practice nurses) during consultations
30 where smoking status is often recorded as part of an assessment of current or
31 future disease risk.
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 Our examination of the age-standardised data suggests that typically an ex-
4 smoker in primary care settings is recorded as a non-smoker when they quit at a
5 young age or had not smoked for a substantial time period. This could be
6 because the patient may not volunteer previous smoking in either initial self-
7 report questionnaire or on questioning by clinicians when it was minor, long ago
8 or they consider it not relevant to their current or future health. It is possible that
9 patients are more reluctant to volunteer ex-smoking habits when data are being
10 held on their medical record and is not anonymous. However, comparing the
11 proportion of individuals with a smoking record in THIN with that of the HSE we
12 found a similar distribution suggesting that most smokers were identified in the
13 first year of their registration in primary care. Similar findings have been
14 observed in the literature by calendar year.[32] While some studies suggest
15 underreporting of smoking among pregnant women in primary care[33] we found
16 no evidence this was a general pattern. With the introduction of the Quality and
17 Outcomes Framework in 2004, there has also been increased incentive to
18 identify smokers in relation to specific disease outcomes.[34, 35] Indeed we
19 found in our previous study that those with respiratory and cardiac conditions
20 were more likely to have any smoking status recorded within the first year of
21 registration.[13] Smoking status was validated in the HSE in 2007 by the use of
22 saliva cotinine samples and was found to be accurate[36].
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52

53 The method of age standardisation then deducing the average time since quitting
54 and reclassifying them to non-smokers in the HSE is relatively crude and
55
56
57
58
59
60

1
2
3 assumes that everyone who becomes an ex-smoker does so at the same time in
4
5 their lives as others in their age group. However, it may be indicative of reporting
6
7 of smoking status at the GP practice, given the results shown in this study.
8
9

10
11
12 An alternative method of dealing with unobserved smoking data is to dichotomise
13
14 smoking status into current smokers and non-current smokers with missing data
15
16 assumed to be non-current smokers. However, it should be noted that this
17
18 solution may be to the detriment of some epidemiological studies where ex-
19
20 smokers who quit recently are at greater risk of disease than non-smokers. For
21
22 example, the 50 year follow up of male British doctors shows that ex-smokers
23
24 had elevated age standardised mortality rates for many diseases.[37, 38]
25
26
27
28
29
30

31
32 Our findings suggest that *in contrast* to health surveys, patients who quit smoking
33
34 at a young age (before 25-30) are likely to be recorded by their general practice
35
36 as a non-smoker instead of an ex-smoker. This has implications for researchers
37
38 using these data sources. To our knowledge this is the first study which seeks to
39
40 deduce and quantify typical time between when a smoker quit and when they are
41
42 no longer perceived as an ex-smoker in primary care. Clinicians, policy-makers
43
44 and researchers who wish to use smoking status in primary care records to
45
46 identify populations at risk of smoking-related diseases can be reassured by our
47
48 findings that using data from new registrations, most current smokers will be
49
50 identified and misclassification of ex-smokers is more likely to have occurred in
51
52 those who have quit smoking at an early age and/ or a long time ago.
53
54
55
56
57
58
59
60

1
2
3 Figure Legends
4
5

6 Figure 1: Smoking status percentages in THIN and the HSE 2008 by age group
7

8 Figure 1 footnotes:

9 Solid line is the Health Survey for England 2008, dashed line is The Health
10 Improvement Network (THIN)
11
12

13
14 Figure 2: Smoking status percentages in THIN and the HSE 2008 by deprivation
15 quintile
16

17 Figure 2 footnotes:

18 *IMD 1 is the least deprived and IMD 5 is the most deprived

19 Darker bars represent the HSE 2008, lighter bars represent THIN

20 Abbreviations: HSE Health Survey for England 2008, IMD Index of multiple
21 deprivation, THIN The Health Improvement Network
22
23

24
25
26 Figure 3: Age group specific percentages of ex-smokers in THIN (after MNAR
27 imputation) and the HSE 2008 (before and after reclassifying ex-smokers in the
28 HSE who quit before the age specified in Table 3 column 3 to be non-smokers)
29

30 Figure 3 footnotes:

31 Abbreviations: THIN The Health Improvement Network, HSE Health Survey for
32 England 2008
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Conflict of interest and funding

The authors have no conflicts of interest to declare. This study is funded by a UK Medical Research Council grant [G0900701]. The funder had no influence over the study design, results or decision to publish this work. JRC was funded by a UK Economic and Social Research Council research fellowship grant [RES-063-27-0257]. IRW was funded by a United Kingdom Medical Research Council grant [U105260558].

Author contributions

LM extracted and analysed the data and wrote the first draft of the paper with help from IP and JRC. KRW and IN provided clinical input and IRW and RWM provided additional statistical input. All authors commented on the paper and helped write subsequent drafts.

Data sharing statement

No data are available

I Dr Louise Marston the Corresponding Author of this article contained within the original manuscript which includes any diagrams & photographs, other illustrative material, video, film or any other material howsoever submitted by the Contributor(s) at any time and related to the Contribution (“the Contribution”) have the right to grant on behalf of all authors and do grant on behalf of all authors, a licence to the BMJ Publishing Group Ltd and its licensees, to permit this Contribution (if accepted) to be published in BMJ Open and any other BMJ Group products and to exploit all subsidiary rights, as set out in the licence at: http://group.bmj.com/products/journals/instructions-for-authors/BMJOpen_licence.pdf

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

References

1 Office for National Statistics Opinions and Lifestyle Survey, Smoking Habits
Amongst Adults, 2012; 2013 (http://www.ons.gov.uk/ons/dcp171776_328041.pdf)
(accessed January 2014)

2 Davies AR, Smeeth L, Grundy EMD. Contribution of changes in incidence and
mortality to trends in the prevalence of coronary heart disease in the UK: 1996-
2005. *Eur Heart J* 2007;2142-2147.

3 Delaney JAC, Daskalopoulou SS, Brophy JM et al Lifestyle variables and the
risk of myocardial infarction in the general practice research database (electronic
article). *BMC Cardiovasc Disord* 2007;38.

4 Douglas IJ, Smeeth L. Exposure to antipsychotics and risk of stroke: self
controlled case series study (electronic article). *Br Med J* 2008;

5 Gelfand JM, Neimann AL, Shin DB et al Risk of myocardial infarction in
patients with psoriasis. *JAMA* 2006;1735-1741.

6 Hippisley-Cox J, Coupland C, Vinogradova Y et al Derivation and validation of
QRISK, a new cardiovascular disease risk score for the United Kingdom:
prospective open cohort study. *Br Med J* 2007;136-141.

1
2
3 7 Osborn DPJ, Levy G, Nazareth I et al Relative risk of cardiovascular and
4 cancer mortality in people with severe mental illness from the United Kingdom's
5 General Practice Research Database. *Arch Gen Psychiatry* 2007;242-249.
6
7
8
9

10
11
12 8 Smeeth L, Thomas SL, Hall AJ et al Risk of myocardial infarction and stroke
13 after acute infection or vaccination. *N Engl J Med* 2004;2611-2618.
14
15
16
17

18
19
20 9 Walters K, Rait G, Petersen I et al Panic disorder and risk of new onset
21 coronary heart disease, acute myocardial infarction, and cardiac mortality: cohort
22 study using the general practice research database. *Eur Heart J* 2008;2981-
23 2988.
24
25
26
27

28
29
30
31 10 Horsfall LJ, Rait G, Walters K et al Serum Bilirubin and Risk of Respiratory
32 Disease and Death *JAMA* 2011;691-697
33
34
35
36

37
38
39 11 Kiri VA, Fabbri LM, Davis KJ et al Inhaled corticosteroids and risk of lung
40 cancer among COPD patients who quit smoking *Respir Med* 2009;85-90
41
42
43
44

45
46 12 Horsfall LJ, Nazareth I, Petersen I. Cardiovascular Events as a Function of
47 Serum Bilirubin Levels in a Large Statin-Treated Cohort. *Circulation* 2012;2556-
48 2564
49
50
51
52

1
2
3 13 Szatkowski L, Lewis S, McNeill A et al Is smoking status routinely recorded
4 when patients register with a new GP? *Fam Pract* 2010;673–675
5
6
7

8
9
10 14 Dhalwani NN, Tata LJ, Coleman T, Fleming KM, Szatkowski L Completeness
11 of Maternal Smoking Status Recording during Pregnancy in United Kingdom
12 Primary Care Data. *PLoS One* 2013;e72218
13
14
15
16

17
18
19 15 Coleman T, Lewis S, Hubbard R et al Impact of contractual financial
20 incentives on the ascertainment and management of smoking in primary care.
21 *Addiction* 2007;102:803e8.
22
23
24
25
26

27
28
29 16 Marston L, Carpenter JR, Walters KR et al Issues in multiple imputation of
30 missing data for large general practice clinical databases *Pharmacoepidemiol*
31 *Drug Saf* 2010;618–626
32
33
34
35
36

37
38
39 17 Weiner MG, Barnhart K, Xie D et al Hormone therapy and coronary heart
40 disease in young women. *Menopause* 2008;86-93.
41
42
43
44

45
46 18 Sterne JAC, White IR, Carlin JB et al. Multiple imputation for missing data in
47 epidemiological and clinical research: potential and pitfalls. *Br Med J* 2009;b2393
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 19 Hippisley-Cox J, Coupland C, Vinogradova Y et al Predicting cardiovascular
4 risk in England and Wales: prospective derivation and validation of QRISK2. *Br*
5
6 *Med J* 2008;1475-1482
7
8
9

10
11
12 20 Collins GS, Altman DG An independent and external validation of QRISK2
13 cardiovascular disease risk score: a prospective open cohort study *Br Med J*
14
15 2010;c2442
16
17
18
19

20
21 21 The Health Improvement Network The Health Improvement Network. London:
22 The Health Improvement Network; 2014 (<http://csdmruk.cegedim.com/>)
23
24 (Accessed January 2014).
25
26
27
28
29

30
31 22 Lis Y, Mann RD The VAMP Research multi-purpose database in the U.K. *J*
32
33 *Clin Epidemiol* 1995;431-443.
34
35
36
37

38 23 Blak BT, Thompson M, Dattani H, Bourke A Generalisability of The Health
39 Improvement Network (THIN) database: demographics, chronic disease
40 prevalence and mortality rates. *Inform Prim Care* 2011;251-255.
41
42
43
44
45

46
47 24 National Centre for Social Research and University College London.
48 Department of Epidemiology and Public Health, *Health Survey for England, 2008*
49 [computer file]. *2nd Edition*. Colchester, Essex: UK Data Archive [distributor],
50
51 October 2010. SN: 6397
52
53
54
55
56
57
58
59
60

1
2
3
4
5
6 25 Noble M, Wright G, Dibben C et al Indices of Deprivation 2004. Report to the
7
8 Office of the Deputy Prime Minister. London: Neighbourhood Renewal Unit; 2004
9
10 (<http://webarchive.nationalarchives.gov.uk/20120919132719/http://www.communi>
11
12 [ties.gov.uk/documents/communities/pdf/131209.pdf](http://www.communities.gov.uk/documents/communities/pdf/131209.pdf)) (Accessed January 2014)
13
14

15
16
17 26 Spratt M, Carpenter J, Sterne JAC et al Strategies for Multiple Imputation in
18
19 Longitudinal Studies *Am. J. Epidemiol.* 2010;478-487
20
21

22
23
24 27 White IR, Royston P, Wood A Multiple imputation using chained equations:
25
26 issues and guidance for practice (tutorial). *Stat Med* 2011;377-399
27
28

29
30
31 28 Graham JW Missing data analysis: Making it work in the real world *Annu Rev*
32
33 *Psychol* 2009;549-576
34
35

36
37
38 29 Stata Corporation. *Stata Statistical Software: Release 11*. College Station, TX:
39
40 Stata Corporation; 2009
41
42

43
44
45 30 Royston P Multiple imputation of missing values: Update of ice. *Stata Journal*
46
47 2005;527-536.
48
49

50
51
52 31 Rubin DB *Multiple imputation for non-response in surveys*. New York: John
53
54 Wiley and Sons; 1987
55
56
57
58
59
60

1
2
3
4
5
6 32 Szatkowski L, Lewis S, McNeill A et al Can data from primary care medical
7 records be used to monitor national smoking prevalence? *J Epidemiol*
8 *Community Health* 2011. doi:10.1136/jech.2010.120154
9
10
11
12
13

14
15 33 Shipton D, Tappin DM, Vadiveloo T et al Reliability of self reported smoking
16 status by pregnant women for estimating smoking prevalence: a retrospective,
17 cross sectional study *Br Med J* 2009;b4347.
18
19
20
21
22
23

24 34 The British Medical Association and NHS Employers Quality and Outcomes
25 Framework guidance for GMS contract 2011/12; 2011
26 ([http://www.nhsemployers.org/Aboutus/Publications/Documents/QOF_guidance_](http://www.nhsemployers.org/Aboutus/Publications/Documents/QOF_guidance_GMS_contract_2011_12.pdf)
27 [GMS_contract_2011_12.pdf](http://www.nhsemployers.org/Aboutus/Publications/Documents/QOF_guidance_GMS_contract_2011_12.pdf)) (Accessed January 2014)
28
29
30
31
32
33

34
35
36 35 Campbell S, Reeves D, Kontopantelis E et al Quality of Primary Care in
37 England with the Introduction of Pay for Performance *N Engl J Med* 2007;181-
38 190
39
40
41
42
43
44

45 36 Wardle H, Mindell J Adult cigarette smoking. In Craig R, Shelton N (Ed.),
46 Health Survey for England 2007. Volume 1. Healthy lifestyles: Knowledge,
47 attitudes and behaviour. 2008;149-176 NHS Information Centre.
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 37 Doll R, Peto R, Boreham J et al Mortality in relation to smoking: 50 years'
4 observations on male British doctors *Br Med J* 2004,
5
6 doi:10.1136/bmj.38142.554479.AE
7
8
9

10
11
12 38 Kenfield SA, Wei EK, Rosner BA et al Burden of smoking on cause-specific
13 mortality: application to the Nurses' Health Study *Tob Control* 2010;248-254
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

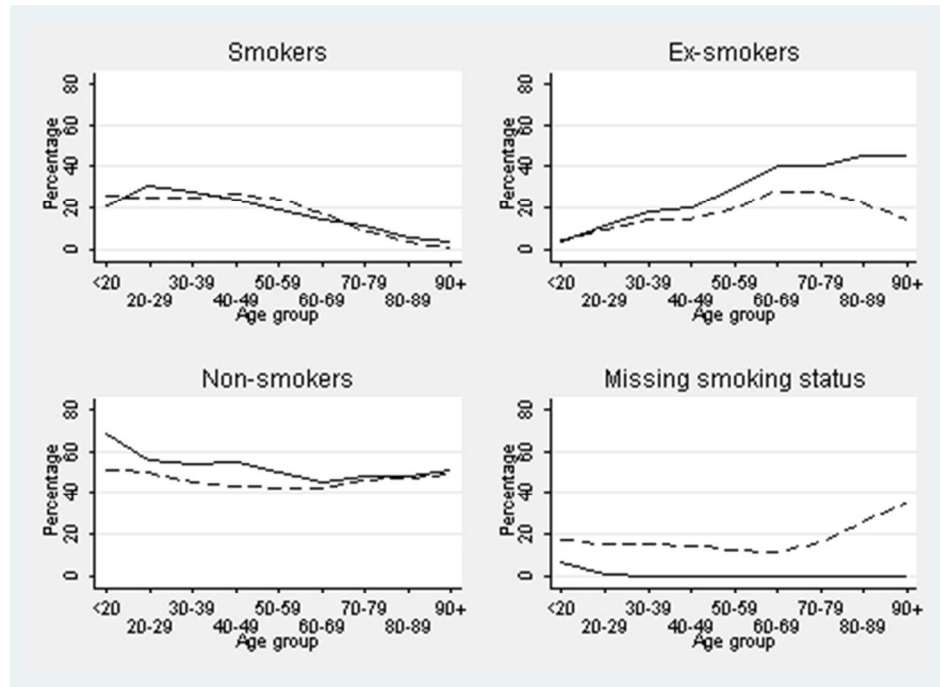


Figure 1: Smoking status percentages in THIN and the HSE 2008 by age group

Solid line is the Health Survey for England 2008, dashed line is The Health Improvement Network (THIN)

166x120mm (72 x 72 DPI)

View only

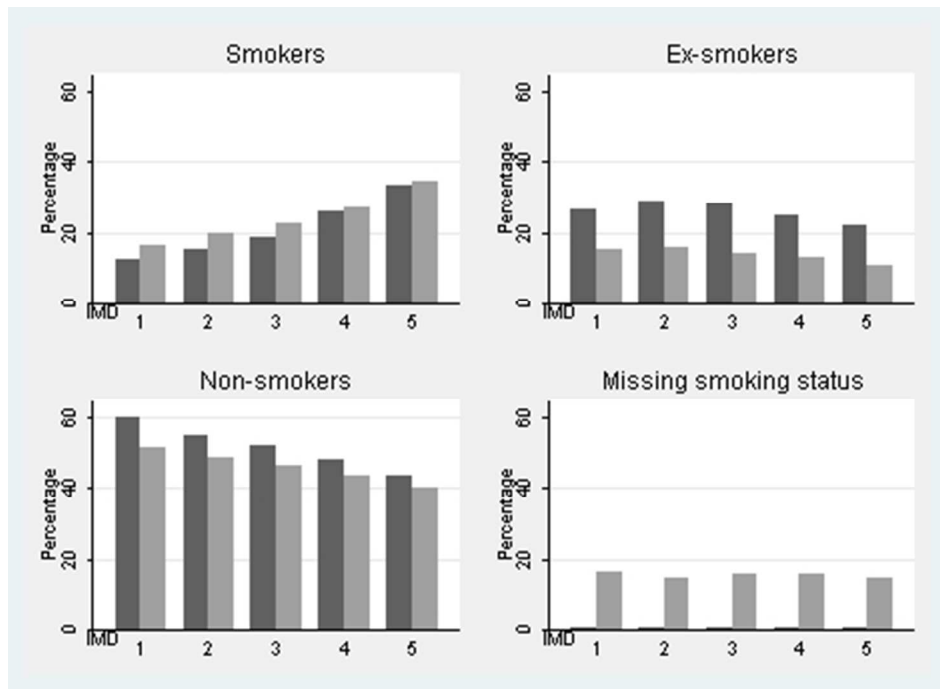


Figure 2: Smoking status percentages in THIN and the HSE 2008 by deprivation quintile

*IMD 1 is the least deprived and IMD 5 is the most deprived

Darker bars represent the HSE 2008, lighter bars represent THIN

Abbreviations: HSE Health Survey for England 2008, IMD Index of multiple deprivation, THIN The Health Improvement Network

166x120mm (72 x 72 DPI)

View Only

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

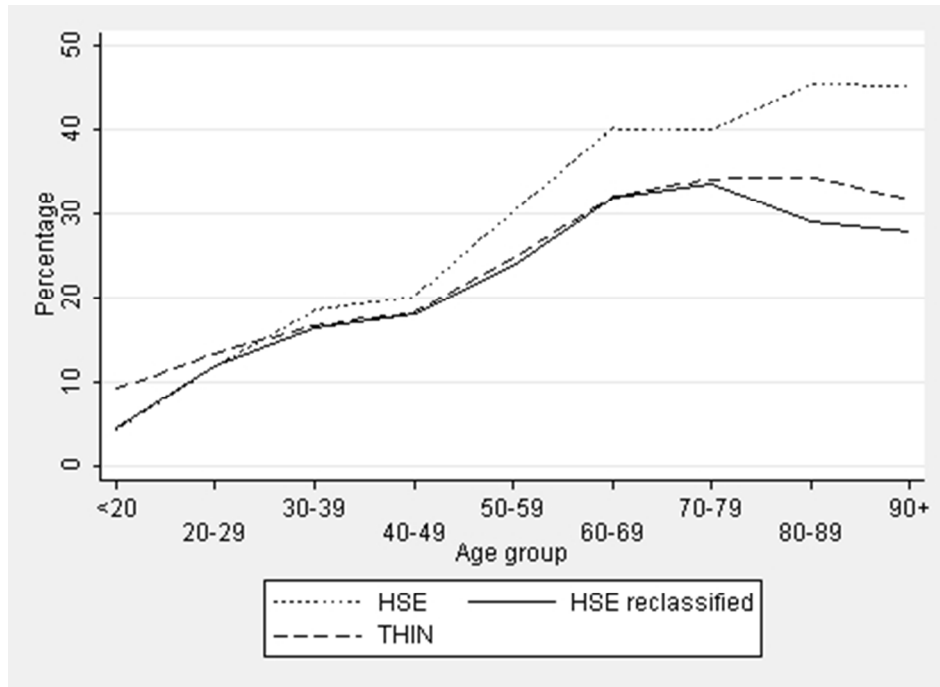


Figure 3: Age group specific percentages of ex-smokers in THIN (after MNAR imputation) and the HSE 2008 (before and after reclassifying ex-smokers in the HSE who quit before the age specified in Table 3 column 3 to be non-smokers)

Abbreviations: THIN The Health Improvement Network, HSE Health Survey for England 2008

166x120mm (72 x 72 DPI)

Review only

STROBE 2007 (v4) Statement—Checklist of items that should be included in reports of *cross-sectional studies*

Section/Topic	Item #	Recommendation	Reported on page #
Title and abstract	1	(a) Indicate the study's design with a commonly used term in the title or the abstract	1
		(b) Provide in the abstract an informative and balanced summary of what was done and what was found	3-4
Introduction			
Background/rationale	2	Explain the scientific background and rationale for the investigation being reported	5-7
Objectives	3	State specific objectives, including any prespecified hypotheses	7
Methods			
Study design	4	Present key elements of study design early in the paper	7-8
Setting	5	Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection	7-8
Participants	6	(a) Give the eligibility criteria, and the sources and methods of selection of participants	8
Variables	7	Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable	8-10
Data sources/ measurement	8*	For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group	8-10
Bias	9	Describe any efforts to address potential sources of bias	8
Study size	10	Explain how the study size was arrived at	NA
Quantitative variables	11	Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen and why	9-10
Statistical methods	12	(a) Describe all statistical methods, including those used to control for confounding	9-11
		(b) Describe any methods used to examine subgroups and interactions	9-11
		(c) Explain how missing data were addressed	9-11
		(d) If applicable, describe analytical methods taking account of sampling strategy	NA
		(e) Describe any sensitivity analyses	NA
Results			

Participants	13*	(a) Report numbers of individuals at each stage of study—eg numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analysed	11
		(b) Give reasons for non-participation at each stage	NA
		(c) Consider use of a flow diagram	NA
Descriptive data	14*	(a) Give characteristics of study participants (eg demographic, clinical, social) and information on exposures and potential confounders	11-12
		(b) Indicate number of participants with missing data for each variable of interest	12
Outcome data	15*	Report numbers of outcome events or summary measures	12
Main results	16	(a) Give unadjusted estimates and, if applicable, confounder-adjusted estimates and their precision (eg, 95% confidence interval). Make clear which confounders were adjusted for and why they were included	12-16
		(b) Report category boundaries when continuous variables were categorized	NA
		(c) If relevant, consider translating estimates of relative risk into absolute risk for a meaningful time period	NA
Other analyses	17	Report other analyses done—eg analyses of subgroups and interactions, and sensitivity analyses	NA
Discussion			
Key results	18	Summarise key results with reference to study objectives	16
Limitations	19	Discuss limitations of the study, taking into account sources of potential bias or imprecision. Discuss both direction and magnitude of any potential bias	17-19
Interpretation	20	Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence	17-18
Generalisability	21	Discuss the generalisability (external validity) of the study results	19
Other information			
Funding	22	Give the source of funding and the role of the funders for the present study and, if applicable, for the original study on which the present article is based	21

*Give information separately for cases and controls in case-control studies and, if applicable, for exposed and unexposed groups in cohort and cross-sectional studies.

Note: An Explanation and Elaboration article discusses each checklist item and gives methodological background and published examples of transparent reporting. The STROBE checklist is best used in conjunction with this article (freely available on the Web sites of PLoS Medicine at <http://www.plosmedicine.org/>, Annals of Internal Medicine at <http://www.annals.org/>, and Epidemiology at <http://www.epidem.com/>). Information on the STROBE Initiative is available at www.strobe-statement.org.

BMJ Open

Smoker, ex-smoker or non-smoker? The validity of routinely recorded smoking status in UK primary care: a cross-sectional study

Journal:	<i>BMJ Open</i>
Manuscript ID:	bmjopen-2014-004958.R1
Article Type:	Research
Date Submitted by the Author:	20-Mar-2014
Complete List of Authors:	Marston, Louise; University College London, Primary Care and Population Health Carpenter, James; London School of Hygiene & Tropical Medicine, Medical Statistics; MRC Clinical Trials Unit, Walters, Kate; University College London, Primary Care and Population Health Morris, Richard; University College London, Primary Care and Population Health Nazareth, Irwin; University College London, Primary Care and Population Health White, Ian; Medical Research Council, Biostatistics Unit Petersen, Irene; University College London, Primary Care and Population Health
Primary Subject Heading:	General practice / Family practice
Secondary Subject Heading:	Smoking and tobacco, Research methods
Keywords:	PRIMARY CARE, STATISTICS & RESEARCH METHODS, EPIDEMIOLOGY

SCHOLARONE™
Manuscripts

1
2
3 **Title: Smoker, ex-smoker or non-smoker? The validity of routinely**
4 **recorded smoking status in UK primary care: a cross-sectional study**
5
6
7
8
9

10 **Authors**

11 Louise Marston¹

12 James R Carpenter^{2, 3}

13 Kate R Walters¹

14 Richard W Morris¹

15 Irwin Nazareth¹,

16 Ian R White⁴

17 Irene Petersen¹

18
19
20
21
22
23
24
25
26
27
28
29
30
31
32 ¹Department of Primary Care and Population Health, University College London,
33 Rowland Hill Street, London, NW3 2PF, United Kingdom

34
35
36 ²Department of Medical Statistics, London School of Hygiene and Tropical
37 Medicine, Keppel Street, London, WC1E 7HT, United Kingdom

38
39
40
41 ³ MRC Clinical Trials Unit, Kingsway, London

42
43
44 ⁴MRC Biostatistics Unit, Cambridge Institute of Public Health, University Forvie
45 Site, Robinson Way, Cambridge, CB2 0SR, United Kingdom
46
47
48

49
50 Please address correspondence to:

51 Dr Louise Marston

52
53
54
55 Department of Primary Care and Population Health
56
57
58
59
60

1
2
3 University College London
4

5 Rowland Hill Street
6

7 London
8

9
10 NW3 2PF
11

12 United Kingdom
13

14
15
16
17
18 l.marston@ucl.ac.uk
19

20 Telephone: +44 20 7794 0500 (36768)
21

22 Fax: +44 20 7794 1224
23

24
25
26 Words: 3,376
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

ABSTRACT

Objectives: To investigate how smoking status is recorded in UK primary care; to evaluate whether appropriate multiple imputation (MI) of smoking status yields results consistent with health surveys.

Setting: UK primary care and a population survey conducted in the community.

Participants: We identified 354,204 patients aged 16 or over in The Health Improvement Network (THIN) primary care database registered with their general practice 2008-2009 and 15,102 individuals aged 16 or over in the Health Survey for England (HSE).

Outcome measures: Age-standardised and age-specific proportions of smokers, ex-smokers and non-smokers in THIN and the HSE before and after multiple imputation (MI). Using information on time since quitting in the HSE, we estimated when ex-smokers are typically recorded as non-smokers in primary care records.

Results: In THIN, smoking status was recorded for 84% of patients within one year of registration. Of these; 28% were smokers (21% in the HSE). After MI of missing smoking data, the proportion of smokers was 25% (missing at random) and 20% (missing not at random). With increasing age, more were identified as ex-smokers in the HSE than THIN. It appears that those who quit before age 30

1
2
3 were less likely to be recorded as an ex-smoker in primary care than people who
4
5 quit later.
6
7
8
9

10 **Conclusions:** Smoking status was relatively well recorded in primary care.
11
12 Misclassification of ex-smokers as non-smokers is likely to occur in those quitting
13
14 smoking at an early age and/ or a long time ago. Those with no smoking status
15
16 information are more likely to be ex or non-smokers than smokers.
17
18
19

20 21 22 **STRENGTHS AND LIMITATIONS OF THIS STUDY** 23

- 24 • This study includes data from 'real' life primary care electronic records
- 25
- 26 • First study to compare the definition of smoking status in primary care with
- 27
- 28 a population survey
- 29
- 30
- 31 • Study focuses on data recorded in the first year after patient registration
- 32
- 33 and may not be applicable to other times.
- 34
- 35
- 36
- 37
- 38

39 **KEYWORDS:** recording of smoking, primary care databases, Health Survey for
40
41 England, missing data, multiple imputation
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

INTRODUCTION

A fifth of the British adult population are smokers [1] and there is still a need for further research into smoking and smoking related diseases including coronary heart disease and stroke, respiratory diseases and cancers. Routinely collected smoking data can be used in clinical practice to identify populations at risk of smoking-related diseases, such as identifying smokers to undergo spirometry testing for early diagnosis of Chronic Obstructive Pulmonary Disease (COPD), or to be invited for smoking cessation services. It is important to understand the accuracy of the data, and whether cases may be missed in those with no recorded smoking status. Electronic health records, including primary care databases, have proved to be very powerful resources for epidemiological and health research.[2-12], allowing research that would be difficult using primary research methods; for example, studying the elderly and people with severe mental illness.[4, 7, 9, 11] Additionally they include millions of patients giving power to study rare conditions. Nevertheless, as they are collected for clinical reasons, they raise a number of issues when used for research; not least of these is missing data.

In order to conduct such research, it is important to understand how smoking status is recorded in primary care and how missing data may be addressed. There is evidence that the recording of smoking status has improved substantially in UK primary care[13, 14] and estimates of *current smoking* are similar to large population surveys.[15, 16] Most general practices now routinely

1
2
3 record smoking status at regular intervals as a part of the Quality Outcome
4 Framework.[17] However, we do not know how the different and non-
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

record smoking status at regular intervals as a part of the Quality Outcome Framework.[17] However, we do not know how the different and non-standardised classifications of *ex and, non smokers* in primary care records compared to the standardised recording of smoking status in population surveys such as the Health Survey for England (HSE).

As noted already, a proportion of patients still lack a smoking status record in their primary care records. It is unclear how to deal with these patients when conducting research where smoking status is either the outcome of the research or an explanatory factor for patients' health.[3, 6, 18, 19] Methodological research has demonstrated that including only patients with complete records can substantially bias the results, especially when the reason for missing data is associated with patient outcomes.[20, 21] In recent years, efforts have been made to address missing data in primary care databases[3, 19, 22] using multiple imputation, though reporting on the comparability of the results of multiple imputation with population data has been sparse. Therefore, it is unclear whether multiple imputation accurately replicates data representing the population.[3, 6, 19, 23] Our previous work on missing data in The Health Improvement Network (THIN) primary care database showed that many health indicator measurements (for example, weight and blood pressure) recorded within the first year of patients' registration at a general practice were comparable with large external datasets before and after multiple imputation.[18] However, smoking status was not directly comparable with data from the Health Survey for

1
2
3 England (HSE). Although the proportion of smokers was similar between THIN
4 and the HSE *before* multiple imputation of data in THIN, the proportion of
5 smokers was substantially higher *after* multiple imputation in THIN. On the other
6 hand, the proportion of ex-smokers was substantially lower in THIN both *before*
7 and *after* imputation compared to the HSE. This suggests that current smokers
8 may be adequately identified using primary care data and most people with
9 missing data on smoking status are likely to be either ex or non-smokers. This
10 has clinical importance as smoking status (including ex-smoking) may be used to
11 identify those at risk of disease, for example chronic obstructive pulmonary
12 disease or cardiovascular disease.
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28

29 In this study we further investigate recording of smoking status in primary care
30 and explore potential reasons for the discrepancy in the proportion of ex-smokers
31 between primary care records and the HSE. Specifically, we seek to deduce
32 when ex-smokers may not be recorded as such in primary care records based on
33 information about time since quitting in the HSE. Finally, we aim to provide a
34 practical solution for imputation of missing smoking status records in routinely
35 collected clinical data.
36
37
38
39
40
41
42
43
44
45
46
47

48 **METHODS**

49 **Study populations**

50 We used data from THIN primary care database, from practices in England that
51 had passed data quality checks, to ensure they were using their computer
52
53
54
55
56
57
58
59
60

1
2
3 system to record all patient consultations.[24-26] In the United Kingdom (UK)
4
5 98% of the population are registered with a National Health Service (NHS)
6
7 general practitioner to receive routine healthcare.[27] THIN is broadly
8
9 representative of all general practices in the UK in terms of age and sex of
10
11 patients, practice size and geographical distribution.[28] The database contains
12
13 information on socio demographics, symptoms, diagnoses, referrals to secondary
14
15 care, prescribing, results of tests and health status indicators. The data provider
16
17 (CSD-MR) obtained overall ethical approval from the South East MREC
18
19 (MREC/03/01/073) and this study was further approved by a THIN scientific
20
21 review committee.
22
23
24
25
26
27
28

29 For this study we selected patients aged 16 years or over who registered with a
30
31 general practice between 1st January 2008 and 31st December 2009
32
33 (N=354,204) and were registered for at least a year. We examined records from
34
35 the first year after the patient registered, hence using data up to the end of 2010.
36
37 Many people have a “new patient check” soon after registration, where
38
39 information on demographics, health indicators and disease status is collected.
40
41
42
43
44
45

46 We compared the distribution of smoking status with that in the HSE from 2008
47
48 for those aged 16 years or over (N=15,102). The HSE is a national annual cross
49
50 sectional interview based survey of approximately 22,000 people.[29] The
51
52 survey includes questions on socio demographics, general health and
53
54 information on smoking status. The HSE has nearly complete records of
55
56
57
58
59
60

1
2
3 smoking (99.3%) and we therefore used the data from patients with complete
4
5 smoking information.
6
7
8
9

10 **Definition of smoking status**

11
12 In THIN, smoking status was recorded by self-report. In many general practices
13
14 this would be on the basis of a questionnaire submitted at the time of registration,
15
16 whereas in other general practices this would be recorded in conjunction with a
17
18 clinical consultation with the general practitioner or practice nurse. GPs and
19
20 nurses may be more interested in the separation between current non-smokers
21
22 and smokers, thus the non-smoking categories may include some people who
23
24 are never smokers as well as some who are ex-smokers in primary care records.
25
26
27

28
29 . In THIN we extracted smoking status data either using Read codes[30] which
30
31 were classified into non-smoker, ex-smoker and smoker with clinical input, or we
32
33 used the categorisation (non-smoker, ex-smoker or current smoker) provided in
34
35 the Additional Health Data. In the HSE, smoking status was defined on the basis
36
37 of a series of questions (see Appendix 1) and individuals who had ever smoked
38
39 (but did not smoke at the time of the interview) would be defined as ex-smokers,
40
41 regardless of their age at quitting and length of time since they quit. The HSE
42
43 holds information on when ex-smokers quit so that age at the time they quit can
44
45 be deduced, whereas this information was not consistently available in THIN.
46
47
48
49
50

51 **Statistical analyses**

52
53
54
55
56
57
58
59
60

1
2
3 Initially, we examined smoking status (smoker, ex-smoker, non-smoker or
4 missing) in THIN and the HSE, overall, by age group, gender and Index of
5 Multiple Deprivation 2004 (IMD) quintile[31]. Then we used multiple imputation
6 to impute missing data in THIN. Multiple imputation via full conditional
7 specification was performed using Stata's "ice" command. [32, 33] Multiple
8 imputation is a statistical method which uses the data available to model the
9 likely distribution of missing data.[20] A number of imputed datasets are
10 produced in each of which plausible values are drawn from the imputation model.
11 The method is designed to correctly reflect the uncertainty surrounding the
12 missing values. With an appropriate imputation model, multiple imputation is an
13 unbiased method of accounting for missing data. It is usually performed under
14 the missing at random (MAR) assumption, but it may also be performed under
15 specific missing not at random (MNAR) assumptions. These methods have been
16 described in greater detail elsewhere.[20, 34-36]

17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39 After preliminary analysis,[34] we included the following variables in the multiple
40 imputation models: age in years, gender and IMD quintile,[31] health indicators:
41 smoking status (three categories, non, ex and current smoker), height, weight,
42 systolic and diastolic blood pressures and disease indicators: type II diabetes,
43 coronary heart disease (CHD) and cerebrovascular accident (CVA). There were
44 missing values for smoking status, blood pressure, weight, height and IMD
45 quintile. . Within the full conditional specification imputation algorithm,
46 continuous variables were imputed using multiple linear regression, smoking
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 status using multinomial regression and IMD quintile using ordered logistic
4 regression. Percentages in each smoking category were obtained using Rubin's
5 Rules.[37] In the first multiple imputation we assumed that smoking data were
6 MAR and hence allowed imputed smoking data of either smokers, non-smokers
7 or ex-smokers (using a MAR assumption; hereafter referred to as MAR MI). In
8 the second multiple imputation we assumed that all smokers had been recorded
9 (so that smoking data were MNAR) and we imputed missing smoking data as
10 either ex-smokers or non-smokers (hereafter referred to as MNAR MI).
11
12
13
14
15
16
17
18
19
20
21
22
23

24 Following multiple imputation we carried out age-specific direct standardisation
25 using the HSE as the standard population and the age-specific proportion in each
26 smoking category from THIN. This was done to account for the fact that the
27 mean age in the HSE was 49 years while the mean age in THIN was 38 years in
28 the year after registration.
29
30
31
32
33
34
35
36
37
38

39 We deduced the average time after which an ex-smoker is no longer classified as
40 an ex-smoker in primary care records by combining information from the HSE on
41 when ex-smokers quit and the age-specific distribution of ex-smokers in THIN
42 after imputation of non and ex-smokers. This was done by ranking the
43 individuals in the HSE in accordance to the length of time since they quit by 10
44 year age groups and then 'reclassifying' individuals who had quit the longest time
45 ago within each age group from ex to non until we reached the same proportion
46 of ex-smokers in the HSE as in THIN. By doing this, we were able to estimate
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 the average time that elapses from quitting smoking after which true ex-smokers
4 are recorded as non-smokers in primary care records.
5
6
7
8
9

10 **RESULTS**

11
12 In total, 354,204 individuals were included from 366 general practices in THIN
13 and 15,102 individuals from the HSE. Individuals in THIN were, on average 11
14 years younger than those in the HSE (38 years versus 49 years, respectively)
15 (Table 1). Smoking status was recorded for 84% in THIN within one year of
16 initial registration. Before multiple imputation of missing data, a greater
17 proportion of people were recorded as smokers in THIN than the HSE (24%
18 versus 21% respectively), and the proportions of ex-smokers and non-smokers
19 differed substantially between THIN and the HSE (Table 1).
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Table 1: Summary statistics for THIN in the first year of registration and the HSE 2008

Variable	THIN		HSE	
	n	%	n	%
Male	164,085	46	6,760	45
Female	190,119	54	8,342	55
Missing sex		0		0
Non-smoker	165,618	47	7,874	52
Ex-smoker	49,874	14	3,966	26
Current smoker	83,526	24	3,158	21
Missing smoking status	55,186	16	104	1
Age years mean (SD)	38	(17)	49	(19)
Missing age		0		0
Least deprived	69,104	20	3,321	22
Quintile 2	71,771	20	3,039	20
Quintile 3	66,422	19	3,010	20
Quintile 4	71,789	20	2,928	19
Most deprived	52,120	15	2,804	19
Missing IMD	22,998	6	0	0

Abbreviations: HSE Health Survey for England 2008; THIN The Health Improvement Network.

Our first analyses used missing as a separate category of smoking, so we refer to those with reported smoking status as “known smokers” and “known ex-smokers”. The proportion of known smokers by age group was similar in THIN and the HSE between 30 and 79 years, but this was not the case for the proportions of known ex-smokers and non-smokers (Figure 1). In the HSE, the proportion of ex-smokers increased from 12% within the 20-29 age group to 46% in the 80-89 age group. In THIN, the proportion of known ex-smokers also increased with age although the overall proportion of known ex-smokers was smaller than in the HSE for all age groups after 20-29 years. Conversely, in the HSE, the proportion of non-smokers decreased slightly from 56% in the 20-29

1
2
3 age group to 48% in the 80-89 age group. Within THIN, the proportion of known
4 non-smokers remained constant with increasing age at around 43%. The
5 proportion of missing smoking data in THIN was relatively constant at less than
6 20% until the 70-79 years age group, but increased substantially thereafter
7 (Figure 1).
8
9
10
11
12
13
14

15
16
17
18 (Figure 1 here)
19
20
21

22 In THIN, the percentage of non-smokers was greater for women (52%) than men
23 (40%) while the percentage of known smokers was smaller for women (21%)
24 than men (27%). There were similar trends in the HSE, although the percentage
25 differences between sexes were smaller (smokers: 22% of men versus 20% of
26 women).
27
28
29
30
31
32
33
34
35

36 The proportions in each smoking status category varied substantially by social
37 deprivation in both THIN and the HSE (Figure 2). In THIN, the percentage of
38 non-smokers decreased from 52% in the least deprived quintile to 40% in the
39 most deprived quintile. The percentage of known ex-smokers decreased slightly
40 with increasing deprivation. In contrast, the percentage of known smokers
41 increased with increasing deprivation from 16% in the least deprived quintile to
42 34% in the most deprived quintile (Figure 2). The patterns were similar in the
43 HSE although the proportion of ex-smokers was substantially larger across all
44 levels of deprivation in the HSE compared to THIN.
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3
4
5
6 (Figure 2 here)
7
8
9

10 **Analyses imputing missing smoking status**

11
12 After MAR MI of THIN, age-standardised smoking prevalences still differed
13 somewhat between THIN and the HSE. For example, 22% were ex-smokers in
14 THIN compared with 26% in the HSE; 25% were smokers in THIN, compared
15 with 21% in the HSE (Table 2).
16
17
18
19
20
21

22
23
24 After MNAR MI of THIN (that is, specifying that missing values are either ex-
25 smokers or non-smokers), the age-standardised prevalence of smoking in THIN
26 was similar to that in the HSE (Table 2). However, the age-specific prevalence of
27 ex-smokers was still greater in the HSE than in THIN. Age-specific analysis
28 showed that this difference was greatest at older ages, and indeed reversed at
29 younger ages. This suggested that individuals who had quit in the less recent
30 past might be classified as non-smokers in THIN but as ex-smokers in
31 HSE.(Figure 3).
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Table 2: Percentages within each smoking status for THIN and the HSE 2008 after various adjustments

Category	THIN			HSE	
	Complete records	After MAR MI. ^{ab}	After MNAR MI ^{ac}	Observed	Reclassifying ex-smokers ^d
	%	%	%	%	%
Non-smoker	55	53	57	53	57
Ex-smoker	17	22	23	26	22
Smoker	28	25	20	21	21

Abbreviations: HSE Health Survey for England 2008; THIN The Health Improvement Network.

^a Directly standardised using the HSE age distribution as standard.

^b Imputed assuming that missing values are smokers, non-smokers or ex-smokers

^c Imputed assuming that missing values are non-smokers or ex-smokers

^d Within each age group, reclassifying the optimum number of ex-smokers as non-smokers based on the distributions shown after MNAR MI.

(Figure 3 here)

The median time since ex-smokers quit in the HSE varied greatly by age group (Table 3), from two years (Interquartile range (IQR): 0, 3) in the under 20s to 40 (IQR: 25, 51) years in those aged 90 or over (Table 3). Equating proportions of ex-smokers in THIN to that in the HSE data suggested the typical time-window after which patients are no longer regarded as ex-smokers in primary care, but instead regarded as non-smokers, varied with age. Thus, typically individuals who registered with a general practice when they were in their forties would no longer be recorded as an ex-smoker if they quit more than 22 years earlier (when they were between 18 and 27 years of age) (Table 3). Individuals registering in their seventies would typically no longer be recorded as ex-smokers if they quit 42 years earlier (when they were between the ages of 28 and 37 years) (Table 3). Yet, most individuals who quit after the age of 30 would still be captured as

ex-smokers when they later registered with a new general practice. Using these age-specific extrapolations to reclassify ex-smokers as non-smokers in the HSE according to when they quit, we can see that the age-specific distributions of ex-smokers in THIN and the reclassified HSE are similar (Figure 3).

Table 3: Age specific centiles of time since quitting smoking in the HSE 2008

Age group	Median time since quitting (years)	Extrapolated number of years since quitting	Extrapolated age when they quit
<20	2	*	*
20-29	3	*	*
30-39	5	14	16 - 25
40-49	10	22	18 - 27
50-59	20	30	20 - 29
60-69	24	35	25 - 34
70-79	30	42	28 - 37
80-89	32	40	40 - 49
90+	40	46	44+

*Not possible to assign an optimal value for reclassification to these age groups
Abbreviations: HSE Health Survey for England 2008

DISCUSSION

The proportion of newly registered patients in THIN between 2008 and 2009 with a record of being a smoker was slightly higher than the HSE in 2008. However, the proportion of individuals recorded as ex-smokers and non-smokers differed substantially between THIN and the HSE. Overall, a larger proportion of individuals were recorded as ex-smokers in the HSE than in THIN and this increased with age. Likewise, the proportion of ex-smokers was substantially larger across all levels of deprivation in the HSE compared to THIN.

1
2
3 Under MAR MI there was a greater percentage of smokers (25%) and a smaller
4 percentage of ex-smokers (22%) in THIN compared with the HSE (smokers 21%,
5 ex-smokers 26%). However, MNAR MI (assuming all missing data were either
6 ex-smokers or non-smokers) slightly increased the proportion of non-smokers
7 (57%) in THIN compared to the HSE (53%), whereas the proportion of ex-
8 smokers (23%) was slightly lower in THIN. Moreover, the latter imputation
9 resulted in a relatively larger percentage of ex-smokers in THIN in those aged
10 under 30 years compared with the HSE. This may be because the imputation
11 model was unable to distinguish between ex and non-smokers in those age
12 groups as both are unlikely to have developed typical later onset diseases which
13 are key predictors of smoking status in the imputation model.
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31

32 There may be several reasons for the discrepancy in the distribution of the
33 smoking categories between THIN and the HSE. In the HSE, the definition of an
34 ex-smoker was highly sensitive and clearly defined.[29] Thus respondents were
35 categorised as ex-smokers even if they were a trivial smoker, smoked for a short
36 period of time and/ or quit many decades ago. Also, the HSE used computer
37 aided personal interviewing; where questions were read to the respondent in a
38 standardised way from the screen and a detailed sequence of questions were
39 asked to ascertain current smoking status. In primary care, while smoking status
40 is systematically recorded in medical records, there is no detailed protocol for
41 recording smoking status and the ascertainment is thus likely to vary by how the
42 information was obtained. Many practices use self-report questionnaires at
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 registration including smoking status. Smoking status is then updated by health
4 professionals (general practitioners and/ or practice nurses) during consultations
5
6 where smoking status is often recorded as part of an assessment of current or
7
8 future disease risk.
9
10
11

12
13
14
15 Our examination of the age-standardised data suggests that typically an ex-
16 smoker in primary care settings is recorded as a non-smoker when they quit at a
17 young age or had not smoked for a substantial time period. This could be
18 because the patient may not volunteer previous smoking in either initial self-
19 report questionnaire or on questioning by clinicians when it was minor, long ago
20 or they consider it not relevant to their current or future health. It is possible that
21 patients are more reluctant to volunteer ex-smoking habits when data are being
22 held on their medical record and is not anonymous. However, comparing the
23 proportion of individuals with a smoking record in THIN with that of the HSE we
24 found a similar distribution suggesting that most smokers were identified in the
25 first year of their registration in primary care. Similar findings have been
26 observed in the literature by calendar year.[18] With the introduction of the
27 Quality and Outcomes Framework in 2004, there has also been increased
28 incentive to identify smokers in relation to specific disease outcomes.[38, 39]
29 Indeed we found in our previous study that those with respiratory and cardiac
30 conditions were more likely to have any smoking status recorded within the first
31 year of registration.[13] Smoking status was validated in the HSE in 2007 by the
32 use of saliva cotinine samples and was found to be accurate[40].
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3
4
5
6 The method of age standardisation then deducing the average time since quitting
7
8 and reclassifying them to non-smokers in the HSE is relatively crude and
9
10 assumes that everyone who becomes an ex-smoker does so at the same time in
11
12 their lives as others in their age group. However, it is likely to be indicative of
13
14 reporting of smoking status at the GP practice, given the results shown in this
15
16 study.
17
18
19
20

21
22 An alternative method of dealing with unobserved smoking data is to dichotomise
23
24 smoking status into current smokers and non-current smokers with missing data
25
26 assumed to be non-current smokers. However, it should be noted that this
27
28 solution may be to the detriment of some epidemiological studies where ex-
29
30 smokers who quit recently are at greater risk of disease than non-smokers. For
31
32 example, the 50 year follow up of male British doctors shows that ex-smokers
33
34 had elevated age standardised mortality rates for many diseases.[41, 42]
35
36
37
38
39
40

41 Our findings suggest that *in contrast* to health surveys, patients who quit smoking
42
43 at a young age (before 25-30) are likely to be recorded by their general practice
44
45 as a non-smoker instead of an ex-smoker. This has implications for researchers
46
47 using these data sources. To our knowledge this is the first study which seeks to
48
49 deduce and quantify typical time between when a smoker quit and when they are
50
51 no longer perceived as an ex-smoker in primary care. Clinicians, policy-makers
52
53 and researchers who wish to use smoking status in primary care records to
54
55
56
57
58
59
60

1
2
3 identify populations at risk of smoking-related diseases can be reassured by our
4
5 findings that using data from new registrations, most current smokers will be
6
7 identified and misclassification of ex-smokers is more likely to have occurred in
8
9 those who have quit smoking at an early age and/ or a long time ago.
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

For peer review only

Conflict of interest and funding

The authors have no conflicts of interest to declare. This study is funded by a UK Medical Research Council grant [G0900701]. The funder had no influence over the study design, results or decision to publish this work. JRC was funded by a UK Economic and Social Research Council research fellowship grant [RES-063-27-0257]. IRW was funded by a United Kingdom Medical Research Council grant [U105260558].

Author contributions

LM extracted and analysed the data and wrote the first draft of the paper with help from IP and JRC. KRW and IN provided clinical input and IRW and RWM provided additional statistical input. All authors commented on the paper and helped write subsequent drafts.

Data sharing statement

No additional data available.

I Dr Louise Marston the Corresponding Author of this article contained within the original manuscript which includes any diagrams & photographs, other illustrative material, video, film or any other material howsoever submitted by the Contributor(s) at any time and related to the Contribution (“the Contribution”) have the right to grant on behalf of all authors and do grant on behalf of all authors, a licence to the BMJ Publishing Group Ltd and its licensees, to permit this Contribution (if accepted) to be published in BMJ Open and any other BMJ Group products and to exploit all subsidiary rights, as set out in the licence at:

http://group.bmj.com/products/journals/instructions-for-authors/BMJOpen_licence.pdf

References

1 Office for National Statistics Opinions and Lifestyle Survey, Smoking Habits Amongst Adults, 2012; 2013 (http://www.ons.gov.uk/ons/dcp171776_328041.pdf) (accessed January 2014)

2 Davies AR, Smeeth L, Grundy EMD. Contribution of changes in incidence and mortality to trends in the prevalence of coronary heart disease in the UK: 1996-2005. *Eur Heart J* 2007;2142-2147.

3 Delaney JAC, Daskalopoulou SS, Brophy JM et al Lifestyle variables and the risk of myocardial infarction in the general practice research database (electronic article). *BMC Cardiovasc Disord* 2007;38.

4 Douglas IJ, Smeeth L. Exposure to antipsychotics and risk of stroke: self controlled case series study (electronic article). *Br Med J* 2008;

5 Gelfand JM, Neimann AL, Shin DB et al Risk of myocardial infarction in patients with psoriasis. *JAMA* 2006;1735-1741.

6 Hippisley-Cox J, Coupland C, Vinogradova Y et al Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: prospective open cohort study. *Br Med J* 2007;136-141.

1
2
3 7 Osborn DPJ, Levy G, Nazareth I et al Relative risk of cardiovascular and
4 cancer mortality in people with severe mental illness from the United Kingdom's
5 General Practice Research Database. *Arch Gen Psychiatry* 2007;242-249.
6
7
8
9

10
11
12 8 Smeeth L, Thomas SL, Hall AJ et al Risk of myocardial infarction and stroke
13 after acute infection or vaccination. *N Engl J Med* 2004;2611-2618.
14
15
16
17

18
19
20 9 Walters K, Rait G, Petersen I et al Panic disorder and risk of new onset
21 coronary heart disease, acute myocardial infarction, and cardiac mortality: cohort
22 study using the general practice research database. *Eur Heart J* 2008;2981-
23 2988.
24
25
26
27

28
29
30
31 10 Horsfall LJ, Rait G, Walters K et al Serum Bilirubin and Risk of Respiratory
32 Disease and Death *JAMA* 2011;691-697
33
34
35
36

37
38
39 11 Kiri VA, Fabbri LM, Davis KJ et al Inhaled corticosteroids and risk of lung
40 cancer among COPD patients who quit smoking *Respir Med* 2009;85-90
41
42
43
44

45
46 12 Horsfall LJ, Nazareth I, Petersen I. Cardiovascular Events as a Function of
47 Serum Bilirubin Levels in a Large Statin-Treated Cohort. *Circulation* 2012;2556-
48 2564
49
50
51
52

1
2
3 13 Szatkowski L, Lewis S, McNeill A et al Is smoking status routinely recorded
4 when patients register with a new GP? *Fam Pract* 2010;673–675
5
6
7

8
9
10 14 Dhalwani NN, Tata LJ, Coleman T, et al. Completeness of Maternal Smoking
11 Status Recording during Pregnancy in United Kingdom Primary Care Data. *PLoS*
12 *One* 2013;e72218
13
14
15

16
17
18 15 Szatkowski L, Lewis S, McNeill A et al Can data from primary care medical
19 records be used to monitor national smoking prevalence? *J Epidemiol*
20 *Community Health* 2012;791-795.
21
22
23
24
25

26
27
28 16 Langley TE, Szatkowski LC, Wythe S, et al Can primary care data be used to
29 monitor regional smoking prevalence? An analysis of The Health Improvement
30 Network primary care data *BMC Public Health* 2011, 11:773
31
32
33
34
35

36
37
38 17 Coleman T, Lewis S, Hubbard R et al Impact of contractual financial
39 incentives on the ascertainment and management of smoking in primary care.
40 *Addiction* 2007;102:803e8.
41
42
43
44
45

46
47
48 18 Marston L, Carpenter JR, Walters KR et al Issues in multiple imputation of
49 missing data for large general practice clinical databases *Pharmacoepidemiol*
50 *Drug Saf* 2010;618–626
51
52
53
54
55

1
2
3 19 Weiner MG, Barnhart K, Xie D et al Hormone therapy and coronary heart
4 disease in young women. *Menopause* 2008;86-93.
5
6
7

8
9
10 20 Sterne JAC, White IR, Carlin JB et al. Multiple imputation for missing data in
11 epidemiological and clinical research: potential and pitfalls. *Br Med J* 2009;b2393
12
13
14

15
16
17 21 Carpenter J, Kenward M Multiple imputation and its application 2013 Wiley
18
19

20
21
22 22 Hippisley-Cox J, Coupland C, Vinogradova Y et al Predicting cardiovascular
23 risk in England and Wales: prospective derivation and validation of QRISK2. *Br*
24 *Med J* 2008;1475-1482
25
26
27

28
29
30
31 23 Collins GS, Altman DG An independent and external validation of QRISK2
32 cardiovascular disease risk score: a prospective open cohort study *Br Med J*
33 2010;c2442
34
35
36
37

38
39
40
41 24 The Health Improvement Network The Health Improvement Network. London:
42 The Health Improvement Network; 2014 (<http://csdmruk.cegedim.com/>)
43
44
45
46 (Accessed January 2014).
47

48
49
50 25 Horsfall L, Walters K, Petersen I. [Identifying periods of acceptable computer](#)
51 [usage in primary care research databases](#). *Pharmacoepidemiol Drug Saf* 2013;
52
53
54
55 64-69
56
57
58
59
60

1
2
3
4
5
6 26 Maguire A, Blak BT, Thompson M. [The importance of defining periods of](#)
7 [complete mortality reporting for research using automated data from primary](#)
8 [care](#). *Pharmacoepidem Drug Safe* 2009; 76-83.
9
10

11
12
13
14
15 27 Lis Y, Mann RD The VAMP Research multi-purpose database in the U.K. *J*
16 *Clin Epidemiol* 1995;431-443.
17
18

19
20
21
22 28 Blak BT, Thompson M, Dattani H, et al Generalisability of The Health
23 Improvement Network (THIN) database: demographics, chronic disease
24 prevalence and mortality rates. *Inform Prim Care* 2011;251-255.
25
26
27

28
29
30
31 29 National Centre for Social Research and University College London.
32 Department of Epidemiology and Public Health, *Health Survey for England, 2008*
33 [computer file]. *2nd Edition*. Colchester, Essex: UK Data Archive [distributor],
34 October 2010. SN: 6397
35
36
37
38
39

40
41
42
43 30 Booth N What are the Read Codes? *Health Libr Rev* 1994;177-82
44
45
46
47

48 31 Noble M, Wright G, Dibben C et al Indices of Deprivation 2004. Report to the
49 Office of the Deputy Prime Minister. London: Neighbourhood Renewal Unit; 2004
50 (<http://webarchive.nationalarchives.gov.uk/20120919132719/http://www.communi>
51 [ties.gov.uk/documents/communities/pdf/131209.pdf](http://www.communities.gov.uk/documents/communities/pdf/131209.pdf)) (Accessed January 2014)
52
53
54
55
56
57
58
59
60

1
2
3
4
5
6 32 Stata Corporation. *Stata Statistical Software: Release 11*. College Station, TX:
7
8 Stata Corporation; 2009
9

10
11
12 33 Royston P Multiple imputation of missing values: Update of ice. *Stata Journal*
13 2005;527-536.
14
15
16

17
18
19
20 34 Spratt M, Carpenter J, Sterne JAC et al Strategies for Multiple Imputation in
21
22 Longitudinal Studies *Am. J. Epidemiol.* 2010;478-487
23
24

25
26
27 35 White IR, Royston P, Wood A Multiple imputation using chained equations:
28
29 issues and guidance for practice (tutorial). *Stat Med* 2011;377-399
30
31

32
33
34 36 Graham JW Missing data analysis: Making it work in the real world *Annu Rev*
35
36 *Psychol* 2009;549-576
37

38
39 37 Rubin DB *Multiple imputation for non-response in surveys*. New York: John
40
41 Wiley and Sons; 1987
42
43

44
45
46 38 The British Medical Association and NHS Employers Quality and Outcomes
47
48 Framework guidance for GMS contract 2011/12; 2011
49
50 ([http://www.nhsemployers.org/Aboutus/Publications/Documents/QOF_guidance_](http://www.nhsemployers.org/Aboutus/Publications/Documents/QOF_guidance_GMS_contract_2011_12.pdf)
51
52 [GMS_contract_2011_12.pdf](http://www.nhsemployers.org/Aboutus/Publications/Documents/QOF_guidance_GMS_contract_2011_12.pdf)) (Accessed January 2014)
53
54
55
56
57
58
59
60

1
2
3 39 Campbell S, Reeves D, Kontopantelis E et al Quality of Primary Care in
4
5
6 England with the Introduction of Pay for Performance *N Engl J Med* 2007;181-
7
8 190
9

10
11
12 40 Wardle H, Mindell J Adult cigarette smoking. In Craig R, Shelton N (Ed.),
13
14 Health Survey for England 2007. Volume 1. Healthy lifestyles: Knowledge,
15
16 attitudes and behaviour. 2008;149-176 NHS Information Centre.
17
18
19

20
21
22 41 Doll R, Peto R, Boreham J et al Mortality in relation to smoking: 50 years'
23
24 observations on male British doctors *Br Med J* 2004,
25
26 doi:10.1136/bmj.38142.554479.AE
27
28
29

30
31
32 42 Kenfield SA, Wei EK, Rosner BA et al Burden of smoking on cause-specific
33
34 mortality: application to the Nurses' Health Study *Tob Control* 2010;248-254
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 Figure Legends
4
5

6 Figure 1: Smoking status percentages in THIN and the HSE 2008 by age group
7

8 Figure 1 footnotes:

9 Solid line is the Health Survey for England 2008, dashed line is The Health
10 Improvement Network (THIN)
11
12

13
14 Figure 2: Smoking status percentages in THIN and the HSE 2008 by deprivation
15 quintile
16

17 Figure 2 footnotes:

18 *IMD 1 is the least deprived and IMD 5 is the most deprived

19 Darker bars represent the HSE 2008, lighter bars represent THIN

20 Abbreviations: HSE Health Survey for England 2008, IMD Index of multiple
21 deprivation, THIN The Health Improvement Network
22
23

24
25
26 Figure 3: Age group specific percentages of ex-smokers in THIN (after MNAR
27 imputation) and the HSE 2008 (before and after reclassifying ex-smokers in the
28 HSE who quit before the age specified in Table 3 column 3 to be non-smokers)
29

30 Figure 3 footnotes:

31 Abbreviations: THIN The Health Improvement Network, HSE Health Survey for
32 England 2008
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3
4
5
6
7
8 **Title: Smoker, ex-smoker or non-smoker? The validity of routinely**
9 **recorded smoking status in UK primary care: a cross-sectional study**
10
11

12
13
14 **Authors**

15 Louise Marston¹

16 James R Carpenter^{2,3}

17 Kate R Walters¹

18 Richard W Morris¹

19 Irwin Nazareth¹

20 Ian R White⁴

21 Irene Petersen¹

22
23
24
25
26
27
28
29
30
31
32 ¹Department of Primary Care and Population Health, University College London,
33 Rowland Hill Street, London, NW3 2PF, United Kingdom

34
35 ²Department of Medical Statistics, London School of Hygiene and Tropical
36 Medicine, Keppel Street, London, WC1E 7HT, United Kingdom

37
38
39 ³ MRC Clinical Trials Unit, Kingsway, London

40
41 ⁴MRC Biostatistics Unit, Cambridge Institute of Public Health, University Forvie
42 Site, Robinson Way, Cambridge, CB2 0SR, United Kingdom
43
44

45
46
47 Please address correspondence to:

48 Dr Louise Marston

49 Department of Primary Care and Population Health
50
51
52
53
54

1
2
3
4
5
6
7
8 University College London

9
10 Rowland Hill Street

11
12 London

13
14 NW3 2PF

15
16 United Kingdom

17
18
19
20 l.marston@ucl.ac.uk

21
22 Telephone: +44 20 7794 0500 (36768)

23
24 Fax: +44 20 7794 1224

25
26
27 | Words: 3, ~~166~~376

ABSTRACT

Objectives: To investigate how smoking status is recorded in UK primary care; to evaluate ~~if~~ whether appropriate multiple imputation (MI) of smoking status yields results consistent with health surveys.

Setting: UK primary care and a population survey conducted in the community.

Participants: We identified 354,204 patients aged 16 or over in The Health Improvement Network (THIN) primary care database registered with their general practice 2008-2009 and 15,102 individuals aged 16 or over in the Health Survey for England (HSE).

Outcome measures: Age-standardised and, age-specific proportions of smokers, ex-smokers and non-smokers in THIN and the HSE before and after multiple imputation (MI). Using information on time since quitting in the HSE, we ~~extrapolated—estimated~~ when ex-smokers ~~may be considered~~ are typically recorded as non-smokers in primary care records.

Results: In THIN, smoking status was recorded for 84% of patients within one year of registration. Of these; 28% were smokers (21% in the HSE). After MI of missing smoking data, the proportion of smokers was 25% (missing at random) and 20% (missing not at random). With increasing age, more were identified as ex-smokers in the HSE than THIN. It appears that those who quit before the ages

1
2
3
4
5
6
7
8
9 | ~~of 25-30 years~~before age 30 were less likely to be recorded as an ex-smoker in
10 primary care than people who quit later.
11

12
13
14 | **Conclusions:** Smoking status ~~is~~was relatively well recorded in primary care.
15 Misclassification of ex-smokers as non-smokers is likely to occur in those quitting
16 smoking at an early age and/ or a long time ago. Those with no smoking status
17 information are more likely to be ex or non-smokers ~~versus~~than smokers.
18
19
20 |
21
22
23

24 **STRENGTHS AND LIMITATIONS OF THIS STUDY**

- 25 • This study includes data from 'real' life primary care electronic records
- 26 • First study to compare the definition of smoking status in primary care
- 27 ~~versus~~with a population survey
- 28 • Study focuses on data recorded in the first year after patient registration
- 29 and may not be applicable to other times.
- 30
31
32
33
34
35
36

37 **KEYWORDS:** recording of smoking, primary care databases, Health Survey for
38 England, missing data, multiple imputation
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54

INTRODUCTION

A fifth of the British adult population are smokers [1] and there is still a need for further research into smoking and smoking related diseases including coronary heart disease and stroke, respiratory diseases and cancers. Routinely collected smoking data can be used in clinical practice to identify populations at risk of smoking-related diseases, such as identifying smokers to ~~have~~ undergo spirometry testing ~~to identify~~ detect these with for early diagnosis of Chronic Obstructive Pulmonary Disease (COPD), or to be invited for smoking cessation services. It is important to understand the accuracy of the data, and whether cases may be missed in those with no recorded smoking status. Electronic health records, including primary care databases, have proved to be very powerful resources for epidemiological and health research.[2-12], ~~—Electronic health records also allow~~ research that would be difficult to capture using primary research methods; for example, studying the elderly and people with severe mental illness.[4, 7, 9, 11] Additionally they include millions of patients giving power to study rare conditions. Nevertheless, as they are collected for clinical reasons, they raise a number of issues when used for research; not least of these is missing data.

In order to conduct such research, it is important to understand how smoking status is recorded in primary care and how missing data may be addressed.

There is evidence that the recording of smoking status has improved substantially in UK primary care[13, 14] and estimates of *current smoking* are

Formatted: Font: Italic

1
2
3
4
5
6
7
8 similar to large population surveys.^[15, 16] ~~and in~~ Most general practices now
9 routinely record smoking status at regular intervals as a part of the Quality
10 Outcome Framework.^[4517] However, we do not know how the different and
11 non-standardised classifications of ~~ex and, non and current~~ smokers in primary
12 care records compared to the standardised recording of smoking status in
13 population surveys such as the Health Survey for England (HSE).
14
15
16
17
18
19
20

Formatted: Font: Italic

21 ~~In addition~~ As noted already, a proportion of patients still lack a smoking status
22 record in their primary care records. It is unclear how to deal with these patients
23 when conducting research where smoking status is either the outcome of the
24 research or an explanatory factor for patients' health.^[3, 6, 4618, 4719]
25 Methodological research has demonstrated that including only patients with
26 complete records can substantially bias the results, especially when the reason
27 for missing data is associated with patient outcomes.^[4820, 21] In recent years,
28 efforts have been made to address missing data in primary care databases^{[3, 47,}
29 ^{19, 22]} using multiple imputation, though reporting on the comparability of the
30 results of multiple imputation with population data has been sparse. Therefore, it
31 is unclear whether multiple imputation accurately replicates data representing the
32 population.^[3, 6, 4719, 2023] Our previous work on missing data in The Health
33 Improvement Network (THIN) primary care database showed that many health
34 indicator measurements (for example, weight and blood pressure) recorded
35 within the first year of patients' registration at a general practice were comparable
36 with large external datasets before and after multiple imputation.^[4618] However,
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52

1
2
3
4
5
6
7
8 smoking status was not directly comparable with data from the Health Survey for
9 England (HSE). Although the proportion of smokers was similar between THIN
10 and the HSE *before* multiple imputation of data in THIN, the proportion of
11 smokers was substantially higher *after* multiple imputation in THIN. On the other
12 hand, the proportion of ex-smokers was substantially lower in THIN both *before*
13 and *after* imputation compared to the HSE. This suggests that current smokers
14 may be adequately identified using primary care data and most people with
15 missing data on smoking status are likely to be either ex or non-smokers. This
16 has clinical importance as smoking status (including ex-smoking) may be used to
17 identify those at risk of disease, for example chronic obstructive pulmonary
18 disease or cardiovascular disease.
19
20
21
22
23
24
25
26
27
28
29
30

31 In this study we further investigate recording of smoking status in primary care
32 and explore potential reasons for the discrepancy in the proportion of ex-smokers
33 between primary care records and the HSE. Specifically, we seek to deduce
34 when ex-smokers may not be recorded as such in primary care records based on
35 information about time since quitting in the HSE. Finally, we aim to provide a
36 practical solution for imputation of missing smoking status records in routinely
37 collected clinical data.
38
39
40
41
42
43
44
45

46 **METHODS**

47 **Study populations**

1
2
3
4
5
6
7
8 We used data from THIN primary care database, from practices in England that
9 had passed data quality checks, to ensure they were using their computer
10 system to record all patient consultations.^[2424-26] In the United Kingdom (UK)
11
12 98% of the population are registered with a National Health Service (NHS)
13
14 general practitioner to receive routine healthcare.^[2227] THIN is broadly
15
16 representative of all general practices in the UK in terms of age and sex of
17
18 patients, practice size and geographical distribution.^[2328] The database
19
20 contains information on socio demographics, symptoms, diagnoses, referrals to
21
22 secondary care, prescribing, results of tests and health status indicators. The
23
24 data provider (CSD-MR) obtained overall ethical approval from the South East
25
26 MREC (MREC/03/01/073) and this study was further approved by a THIN
27
28 scientific review committee.
29
30
31
32

33 For this study we selected patients aged 16 years or over who registered with a
34
35 general practice between 1st January 2008 and 31st December 2009
36
37 (N=354,204) and were registered for at least a year. ~~w~~We examined records
38
39 from the first year after the patient registered, hence using data up to the end of
40
41 2010. Many people have a “new patient check” soon after registration, where
42
43 information on demographics, health indicators and disease status is collected.
44
45
46

47 We compared the distribution of smoking status with that in the HSE from 2008
48
49 for those aged 16 years or over (N=15,102). The HSE is a national annual cross
50
51 sectional interview based survey of approximately 22,000 people.^[2429] The
52
53
54

1
2
3
4
5
6
7
8 survey includes questions on socio demographics, general health and
9 information on smoking status. The HSE has nearly complete records of
10 smoking (99.3%) and we therefore used the data from patients with complete
11 smoking information.
12
13
14

15 16 17 18 **Definition of smoking status**

19 In THIN, smoking status was recorded by self-report. In many general practices
20 this would be on the basis of a questionnaire submitted at the time of registration,
21 whereas in other general practices this would be recorded in conjunction with a
22
23
24

25
26 GPs and
27 nurses may be more interested in the separation between current non-smokers
28 and smokers, thus the non-smoking categories may include some people who
29 are never smokers as well as some who are ex-smokers in primary care records.
30
31 ~~Patients would be classed as current non-smoker, or current smokers. In some~~
32 ~~instance the non-smokers would be classified as ex-smokers but this was~~
33 ~~variably defined from one practice to another. In THIN we extracted smoking~~
34 ~~status data either using Read codes[30] which were classified into non-smoker,~~
35 ~~ex-smoker and smoker with clinical input, or we used the categorisation (non-~~
36 ~~smoker, ex-smoker or current smoker) provided in the Additional Health Data. In~~
37
38
39
40
41
42
43
44

45 the HSE, smoking status was defined on the basis of a series of questions (see
46 Appendix 1) and individuals who had ever smoked (but did not smoke at the time
47 of the interview) would be defined as ex-smokers, regardless of their age at
48 quitting and length of time since they quit. The HSE holds information on when
49
50
51
52
53
54

1
2
3
4
5
6
7
8 ex-smokers quit so that age at the time they quit can be deduced, whereas this
9 information was not consistently available in THIN.
10

11 12 13 **Statistical analyses**

14
15 Initially, we examined smoking status (smoker, ex-smoker, non-smoker or
16 missing) in THIN and the HSE, overall, by age group, gender and Index of
17 Multiple Deprivation 2004 (IMD) quintile[2531]. Then we used multiple
18 imputation to impute missing ~~smoking status data~~ in THIN. Multiple imputation
19 with chained equations via full conditional specification was performed using
20 Stata's "ice" command. [32, 33] Multiple imputation is a statistical method which
21 uses the data available to model the likely distribution of missing data.[4820] A
22 number of imputed datasets are produced in each of which plausible values are
23 drawn from the imputation model. The method is designed to correctly reflect the
24 uncertainty surrounding the missing values. With an appropriate imputation
25 model, multiple imputation is an unbiased method of accounting for missing data.
26 It is usually performed under the missing at random (MAR) assumption, but it
27 may also be performed under specific missing not at random (MNAR)
28 assumptions. These methods have been described in greater detail
29 elsewhere.[4820, 26-2834-36]
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45

46
47 After preliminary analysis,[2634] we included the following variables in the
48 multiple imputation models: age in years, gender and IMD quintile,[2531] health
49 indicators: smoking status (three categories, non, ex and current smoker), height,
50
51
52
53
54

1
2
3
4
5
6
7
8 weight, systolic and diastolic blood pressures and disease indicators: type II
9 diabetes, coronary heart disease (CHD) and cerebrovascular accident (CVA).
10
11 ~~There were missing values for smoking status, blood pressure, weight, height~~
12 ~~and IMD quintile. Multiple imputation was performed using Chained Equations~~
13 ~~using the ice command using Stata 11.[29, 30]~~ Within the full conditional
14 specification imputation algorithm, Continuous variables were imputed using
15 multiple linear regression, smoking status using multinomial regression and IMD
16 quintile using ordered logistic regression. Percentages in each smoking category
17 were obtained using Rubin's Rules.^[3437] In the first multiple imputation we
18 assumed that smoking data were MAR and hence allowed imputed smoking data
19 of either smokers, non-smokers or ex-smokers (using a MAR assumption;
20 hereafter referred to as MAR MI). In the second multiple imputation we assumed
21 that all smokers had been recorded (so that smoking data were MNAR) and we
22 imputed missing smoking data as either ex-smokers or non-smokers (hereafter
23 referred to as MNAR MI).
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38

39 Following multiple imputation we carried out age-specific direct standardisation
40 using the HSE as the standard population and the age-specific proportion in each
41 smoking category from THIN. This was done to account for the fact that the
42 mean age in the HSE was 49 years while the mean age in THIN was 38 years in
43 the year after registration.
44
45
46
47
48
49
50
51
52
53
54

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

We deduced the average time after which an ex-smoker is no longer classified as an ex-smoker in primary care records by combining information from the HSE on when ex-smokers quit and the age-specific distribution of ex-smokers in THIN after imputation of non and ex-smokers. This was done by ranking the individuals in the HSE in accordance to the length of time since they quit by 10 year age groups and then 'reclassifying' individuals who had quit the longest time ago within each age group from ex to non until we reached the same proportion of ex-smokers in the HSE as in THIN. [By doing this, we were able to estimate the average time that elapses from quitting smoking after which true ex-smokers are recorded as non-smokers in primary care records.](#)

RESULTS

In total, 354,204 individuals were included from [366 general practices in](#) THIN and 15,102 [individuals](#) from the HSE. Individuals in THIN were, on average 11 years younger than those in the HSE (38 years versus 49 years, respectively) (Table 1). Smoking status was recorded for 84% in THIN within one year of initial registration. Before multiple imputation of missing data, a greater proportion of people were recorded as smokers in THIN than the HSE (24% versus 21% respectively), and the proportions of ex-smokers and non-smokers differed substantially between THIN and the HSE (Table 1).

Table 1: Summary statistics for THIN in the first year of registration and the HSE 2008

Variable	THIN		HSE	
	n	%	n	%
Male	164,085	46	6,760	45
Female	190,119	54	8,342	55
Missing sex		0		0
Non-smoker	165,618	47	7,874	52
Ex-smoker	49,874	14	3,966	26
Current smoker	83,526	24	3,158	21
Missing smoking status	55,186	16	104	1
Age years mean (SD)	38	(17)	49	(19)
Missing age		0		0*
<u>Least deprived</u>	<u>69,104</u>	<u>20</u>	<u>3,321</u>	<u>22</u>
<u>Quintile 2</u>	<u>71,771</u>	<u>20</u>	<u>3,039</u>	<u>20</u>
<u>Quintile 3</u>	<u>66,422</u>	<u>19</u>	<u>3,010</u>	<u>20</u>
<u>Quintile 4</u>	<u>71,789</u>	<u>20</u>	<u>2,928</u>	<u>19</u>
<u>Most deprived</u>	<u>52,120</u>	<u>15</u>	<u>2,804</u>	<u>19</u>
<u>Missing IMD</u>	<u>22,998</u>	<u>6</u>	<u>0</u>	<u>0</u>

Abbreviations: HSE Health Survey for England 2008; THIN The Health Improvement Network.

Our first analyses used missing as a separate category of smoking, so we refer to those with reported smoking status as “known smokers” and “known ex-smokers”. The proportion of known smokers by age group was similar in THIN and the HSE between 30 and 79 years, but this was not the case for the proportions of known ex-smokers and non-smokers (Figure 1). In the HSE, the proportion of ex-smokers increased from 12% within the 20-29 age group to 46% in the 80-89 age group. In THIN, the proportion of known ex-smokers also increased with age although the overall proportion of known ex-smokers was smaller than in the HSE for all age groups after 20-29 years. Conversely, in the HSE, the proportion of non-smokers decreased slightly from 56% in the 20-29

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

age group to 48% in the 80-89 age group. Within THIN, the proportion of known non-smokers remained constant with increasing age at around 43%. The proportion of missing smoking data in THIN was relatively constant at less than 20% until the 70-79 years age group, but increased substantially thereafter (Figure 1).

(Figure 1 here)

In THIN, the percentage of non-smokers was greater for women (52%) than men (40%) while the percentage of known smokers was smaller for women (21%) than men (27%). There were similar trends in the HSE, although the percentage differences between sexes were smaller (smokers: 22% of men versus 20% of women).

The proportions in each smoking status category varied substantially by social deprivation in both THIN and the HSE (Figure 2). In THIN, the percentage of non-smokers decreased from 52% in the least deprived quintile to 40% in the most deprived quintile. The percentage of known ex-smokers decreased slightly with increasing deprivation. In contrast, the percentage of known smokers increased with increasing deprivation from 16% in the least deprived quintile to 34% in the most deprived quintile (Figure 2). The patterns were similar in the HSE although the proportion of ex-smokers was substantially larger across all levels of deprivation in the HSE compared to THIN.

1
2
3
4
5
6
7
8
9
10 (Figure 2 here)
11
12
13

14 **Analyses imputing missing smoking status**

15
16 After MAR MI of THIN, age-standardised smoking prevalences still differed
17 somewhat between THIN and the HSE. For example, 22% were ex-smokers in
18 THIN compared with 26% in the HSE; 25% were smokers in THIN, compared
19 with 21% in the HSE (Table 2).
20
21
22
23

24
25
26 After MNAR MI of THIN (that is, regarding specifying that missing values as are
27 either ex-smokers or non-smokers), the age-standardised prevalence of smoking
28 in THIN was similar to that in the HSE (Table 2). However, the age-specific
29 prevalence of ex-smokers was still greater in the HSE than in THIN. Age-specific
30 analysis showed that this difference was greatest at older ages, and indeed
31 reversed at younger ages. This suggested that individuals who had quit in the
32 less recent past might be classified as non-smokers in THIN but as ex-smokers
33 in HSE.(Figure 3).
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Table 2: Percentages within each smoking status for THIN and the HSE 2008 after various adjustments

Category	THIN			HSE	
	Complete caserecords	After MAR MI ^{ab}	After MNAR MI ^{ac}	Observed	Reclassifying ex-smokers ^d
	%	%	%	%	%
Non-smoker	55	53	57	53	57
Ex-smoker	17	22	23	26	22
Smoker	28	25	20	21	21

Abbreviations: HSE Health Survey for England 2008; THIN The Health Improvement Network.

^a Directly standardised using the HSE age distribution as standard.

^b Imputed assuming that missing values are smokers, non-smokers or ex-smokers

^c Imputed assuming that missing values are non-smokers or ex-smokers

^d Within each age group, reclassifying the optimum number of ex-smokers as non-smokers based on the distributions shown after MNAR MI.

(Figure 3 here)

The median time since ex-smokers quit in the HSE varied greatly by age group (Table 3), from two years (Interquartile range (IQR): 0, 3) in the under 20s to 40 (IQR: 25, 51) years in those aged 90 or over (Table 3). Equating proportions of ex-smokers in THIN to that in the HSE data suggested the typical time-window after which patients are no longer regarded as ex-smokers in primary care, but instead regarded as non-smokers, varied with age. Thus, typically individuals who registered with a general practice when they were in their forties would no longer be recorded as an ex-smoker if they quit more than 22 years earlier (when they were between 18 and 27 years of age) (Table 3). Individuals registering in their seventies would typically no longer be recorded as ex-smokers if they quit 42 years earlier (when they were between the ages of 28 and 37 years) (Table 3). Yet, most individuals who quit after the age of 30 would still be captured as

ex-smokers when they later registered with a new general practice. Using these age-specific extrapolations to reclassify ex-smokers as non-smokers in the HSE according to when they quit, we can see that the age-specific distributions of ex-smokers in THIN and the reclassified HSE are similar (Figure 3).

Table 3: Age specific centiles of time since quitting smoking in the HSE 2008

Age group	Median time since quitting (years)	Extrapolated number of years since quitting	Extrapolated age when they quit
<20	2	*	*
20-29	3	*	*
30-39	5	14	16 - 25
40-49	10	22	18 - 27
50-59	20	30	20 - 29
60-69	24	35	25 - 34
70-79	30	42	28 - 37
80-89	32	40	40 - 49
90+	40	46	44+

*Not possible to assign an optimal value for reclassification to these age groups
Abbreviations: HSE Health Survey for England 2008

DISCUSSION

The proportion of newly registered patients in THIN between 2008 and 2009 with a record of being a smoker was slightly higher than the HSE in 2008. However, the proportion of individuals recorded as ex-smokers and non-smokers differed substantially between THIN and the HSE. Overall, a larger proportion of individuals were recorded as ex-smokers in the HSE than in THIN and this increased with age. Likewise, the proportion of ex-smokers was substantially larger across all levels of deprivation in the HSE compared to THIN.

1
2
3
4
5
6
7
8 Under MAR MI there was a greater percentage of smokers (25%) and a smaller
9 percentage of ex-smokers (22%) in THIN compared with the HSE (smokers 21%,
10 ex-smokers 26%). However, ~~under~~-MNAR MI (assuming all missing data were
11 either ex-smokers or non-smokers) slightly increased the proportion of non-
12 smokers (57%) in THIN compared to the HSE (53%), whereas the proportion of
13 ex-smokers (23%) was slightly lower in THIN. Moreover, the latter imputation
14 resulted in a relatively larger percentage of ex-smokers in THIN in those aged
15 under 30 years compared with the HSE. This may be because the imputation
16 model was unable to distinguish between ex and non-smokers in those age
17 groups as both are unlikely to have developed typical later onset diseases which
18 are key predictors of smoking status in the imputation model.
19
20
21
22
23
24
25
26
27
28
29
30

31 There may be several reasons for the discrepancy in the distribution of the
32 smoking categories between THIN and the HSE. In the HSE, the definition of an
33 ex-smoker was highly sensitive and clearly defined.^[2429] Thus respondents
34 were categorised as ex-smokers even if they were a trivial smoker, smoked for a
35 short period of time and/ or quit many decades ago. Also, the HSE used
36 computer aided personal interviewing; where questions were read to the
37 respondent in a standardised way from the screen and a detailed sequence of
38 questions were asked to ascertain current smoking status. In primary care, while
39 smoking status is systematically recorded in medical records, there is no detailed
40 protocol for recording smoking status and the ascertainment is thus likely to vary
41 by how the information was obtained. Many practices use self-report
42
43
44
45
46
47
48
49
50
51
52
53
54

1
2
3
4
5
6
7
8 questionnaires at registration including smoking status. Smoking status is then
9 updated by health professionals (general practitioners and/ or practice nurses)
10 during consultations where smoking status is often recorded as part of an
11 assessment of current or future disease risk.
12
13
14

15
16
17
18 Our examination of the age-standardised data suggests that typically an ex-
19 smoker in primary care settings is recorded as a non-smoker when they quit at a
20 young age or had not smoked for a substantial time period. This could be
21 because the patient may not volunteer previous smoking in either initial self-
22 report questionnaire or on questioning by clinicians when it was minor, long ago
23 or they consider it not relevant to their current or future health. It is possible that
24 patients are more reluctant to volunteer ex-smoking habits when data are being
25 held on their medical record and is not anonymous. However, comparing the
26 proportion of individuals with a smoking record in THIN with that of the HSE we
27 found a similar distribution suggesting that most smokers were identified in the
28 first year of their registration in primary care. Similar findings have been
29 observed in the literature by calendar year.^[32,18] ~~While some studies suggest~~
30 ~~underreporting of smoking among pregnant women in primary care~~^[33] ~~we found~~
31 ~~no evidence this was a general pattern.~~—With the introduction of the Quality and
32 Outcomes Framework in 2004, there has also been increased incentive to
33 identify smokers in relation to specific disease outcomes.^[34,38, 35,39] Indeed we
34 found in our previous study that those with respiratory and cardiac conditions
35 were more likely to have any smoking status recorded within the first year of
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3
4
5
6
7
8 registration.[13] Smoking status was validated in the HSE in 2007 by the use of
9
10 saliva cotinine samples and was found to be accurate[3640].
11

12
13
14 The method of age standardisation then deducing the average time since quitting
15 and reclassifying them to non-smokers in the HSE is relatively crude and
16 assumes that everyone who becomes an ex-smoker does so at the same time in
17 their lives as others in their age group. However, it may-is likely to be indicative
18 of reporting of smoking status at the GP practice, given the results shown in this
19 study.
20
21
22
23
24
25

26
27
28 An alternative method of dealing with unobserved smoking data is to dichotomise
29 smoking status into current smokers and non-current smokers with missing data
30 assumed to be non-current smokers. However, it should be noted that this
31 solution may be to the detriment of some epidemiological studies where ex-
32 smokers who quit recently are at greater risk of disease than non-smokers. For
33 example, the 50 year follow up of male British doctors shows that ex-smokers
34 had elevated age standardised mortality rates for many diseases.[37,3841, 42]
35
36
37
38
39
40
41

42
43 Our findings suggest that *in contrast* to health surveys, patients who quit smoking
44 at a young age (before 25-30) are likely to be recorded by their general practice
45 as a non-smoker instead of an ex-smoker. This has implications for researchers
46 using these data sources. To our knowledge this is the first study which seeks to
47 deduce and quantify typical time between when a smoker quit and when they are
48
49
50
51
52
53
54

1
2
3
4
5
6
7
8 no longer perceived as an ex-smoker in primary care. Clinicians, policy-makers
9 and researchers who wish to use smoking status in primary care records to
10 identify populations at risk of smoking-related diseases can be reassured by our
11 findings that using data from new registrations, most current smokers will be
12 identified and misclassification of ex-smokers is more likely to have occurred in
13 those who have quit smoking at an early age and/ or a long time ago.
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54

1
2
3
4
5
6
7
8 Figure Legends
9

10 Figure 1: Smoking status percentages in THIN and the HSE 2008 by age group
11

12 Figure 1 footnotes:

13 Solid line is the Health Survey for England 2008, dashed line is The Health
14 Improvement Network (THIN)
15

16
17 Figure 2: Smoking status percentages in THIN and the HSE 2008 by deprivation
18 quintile
19

20 Figure 2 footnotes:

21 *IMD 1 is the least deprived and IMD 5 is the most deprived

22 Darker bars represent the HSE 2008, lighter bars represent THIN

23 Abbreviations: HSE Health Survey for England 2008, IMD Index of multiple
24 deprivation, THIN The Health Improvement Network
25

26
27 Figure 3: Age group specific percentages of ex-smokers in THIN (after MNAR
28 imputation) and the HSE 2008 (before and after reclassifying ex-smokers in the
29 HSE who quit before the age specified in Table 3 column 3 to be non-smokers)
30

31 Figure 3 footnotes:

32 Abbreviations: THIN The Health Improvement Network, HSE Health Survey for
33 England 2008
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Conflict of interest and funding

The authors have no conflicts of interest to declare. This study is funded by a UK Medical Research Council grant [G0900701]. The funder had no influence over the study design, results or decision to publish this work. JRC was funded by a UK Economic and Social Research Council research fellowship grant [RES-063-27-0257]. IRW was funded by a United Kingdom Medical Research Council grant [U105260558].

Author contributions

LM extracted and analysed the data and wrote the first draft of the paper with help from IP and JRC. KRW and IN provided clinical input and IRW and RWM provided additional statistical input. All authors commented on the paper and helped write subsequent drafts.

Data sharing statement

No data are available

I Dr Louise Marston the Corresponding Author of this article contained within the original manuscript which includes any diagrams & photographs, other illustrative material, video, film or any other material howsoever submitted by the Contributor(s) at any time and related to the Contribution ("the Contribution") have the right to grant on behalf of all authors and do grant on behalf of all authors, a licence to the BMJ Publishing Group Ltd and its licensees, to permit this Contribution (if accepted) to be published in BMJ Open and any other BMJ Group products and to exploit all subsidiary rights, as set out in the licence at: (http://group.bmj.com/products/journals/instructions-for-authors/BMJOpen_licence.pdf)

References

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1 Office for National Statistics Opinions and Lifestyle Survey, Smoking Habits Amongst Adults, 2012; 2013 (http://www.ons.gov.uk/ons/dcp171776_328041.pdf) (accessed January 2014)

2 Davies AR, Smeeth L, Grundy EMD. Contribution of changes in incidence and mortality to trends in the prevalence of coronary heart disease in the UK: 1996-2005. *Eur Heart J* 2007;2142-2147.

3 Delaney JAC, Daskalopoulou SS, Brophy JM et al Lifestyle variables and the risk of myocardial infarction in the general practice research database (electronic article). *BMC Cardiovasc Disord* 2007;38.

4 Douglas IJ, Smeeth L. Exposure to antipsychotics and risk of stroke: self controlled case series study (electronic article). *Br Med J* 2008;

5 Gelfand JM, Neimann AL, Shin DB et al Risk of myocardial infarction in patients with psoriasis. *JAMA* 2006;1735-1741.

6 Hippisley-Cox J, Coupland C, Vinogradova Y et al Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: prospective open cohort study. *Br Med J* 2007;136-141.

1
2
3
4
5
6
7
8 7 Osborn DPJ, Levy G, Nazareth I et al Relative risk of cardiovascular and
9 cancer mortality in people with severe mental illness from the United Kingdom's
10 General Practice Research Database. *Arch Gen Psychiatry* 2007;242-249.
11
12

13
14
15
16 8 Smeeth L, Thomas SL, Hall AJ et al Risk of myocardial infarction and stroke
17 after acute infection or vaccination. *N Engl J Med* 2004;2611-2618.
18
19

20
21
22 9 Walters K, Rait G, Petersen I et al Panic disorder and risk of new onset
23 coronary heart disease, acute myocardial infarction, and cardiac mortality: cohort
24 study using the general practice research database. *Eur Heart J* 2008;2981-
25 2988.
26
27
28

29
30
31 10 Horsfall LJ, Rait G, Walters K et al Serum Bilirubin and Risk of Respiratory
32 Disease and Death *JAMA* 2011;691-697
33
34

35
36
37 11 Kiri VA, Fabbri LM, Davis KJ et al Inhaled corticosteroids and risk of lung
38 cancer among COPD patients who quit smoking *Respir Med* 2009;85-90
39
40
41

42
43 12 Horsfall LJ, Nazareth I, Petersen I. Cardiovascular Events as a Function of
44 Serum Bilirubin Levels in a Large Statin-Treated Cohort. *Circulation* 2012;2556-
45 2564
46
47
48

1
2
3
4
5
6
7
8
9 13 Szatkowski L, Lewis S, McNeill A et al Is smoking status routinely recorded
10 when patients register with a new GP? *Fam Pract* 2010;673–675

11
12
13
14 14 Dhalwani NN, Tata LJ, Coleman T, Fleming KM, Szatkowski L Completeness
15 of Maternal Smoking Status Recording during Pregnancy in United Kingdom
16 Primary Care Data. *PLoS One* 2013;e72218

17
18
19
20
21
22 [15 Szatkowski L, Lewis S, McNeill A et al Can data from primary care medical](#)
23 [records be used to monitor national smoking prevalence? *J Epidemiol*](#)
24 [Community Health](#) 2012;791-795.

25
26
27
28
29
30 [16 Langley TE, Szatkowski LC, Wythe S, et al Can primary care data be used to](#)
31 [monitor regional smoking prevalence? An analysis of The Health Improvement](#)
32 [Network primary care data *BMC Public Health* 2011, 11:773](#)

33
34
35
36
37 ~~15-17~~ Coleman T, Lewis S, Hubbard R et al Impact of contractual financial
38 incentives on the ascertainment and management of smoking in primary care.
39 *Addiction* 2007;102:803e8.

40
41
42
43
44
45 ~~16-18~~ Marston L, Carpenter JR, Walters KR et al Issues in multiple imputation of
46 missing data for large general practice clinical databases *Pharmacoepidemiol*
47 *Drug Saf* 2010;618–626

1
2
3
4
5
6
7
8 | [17-19](#) Weiner MG, Barnhart K, Xie D et al Hormone therapy and coronary heart
9 disease in young women. *Menopause* 2008;86-93.
10
11

12
13
14 | [18-20](#) Sterne JAC, White IR, Carlin JB et al. Multiple imputation for missing data
15 in epidemiological and clinical research: potential and pitfalls. *Br Med J*
16 2009;b2393
17
18
19

20
21
22 | [21](#) Carpenter J, Kenward M. Multiple imputation and its application 2013 Wiley

Formatted: Font: (Default) Arial, 12 pt, Not Bold

23
24
25
26 | [19-22](#) Hippisley-Cox J, Coupland C, Vinogradova Y et al Predicting
27 cardiovascular risk in England and Wales: prospective derivation and validation
28 of QRISK2. *Br Med J* 2008;1475-1482
29
30
31

32
33 | [20-23](#) Collins GS, Altman DG An independent and external validation of QRISK2
34 cardiovascular disease risk score: a prospective open cohort study *Br Med J*
35 2010;c2442
36
37
38

39
40
41 | [21-24](#) The Health Improvement Network The Health Improvement Network.
42 London: The Health Improvement Network; 2014 (<http://csdmruk.cegedim.com/>)
43 (Accessed January 2014).
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3
4
5
6
7
8
9 [25 Horsfall L, Walters K, Petersen I. Identifying periods of acceptable computer
10 usage in primary care research databases. *Pharmacoepidemiol Drug Saf* 2013;
11 \[64-69\]\(#\)](#)

12
13
14
15
16 [26 Maguire A, Blak BT, Thompson M. The importance of defining periods of
17 complete mortality reporting for research using automated data from primary
18 care. *Pharmacoepidem Drug Safe* 2009; 76-83.](#)

19
20
21
22
23 [22-27](#) Lis Y, Mann RD The VAMP Research multi-purpose database in the U.K. *J
24 Clin Epidemiol* 1995;431-443.

25
26
27
28
29 [23-28](#) Blak BT, Thompson M, Dattani H, ~~Bourke Aet al~~ Generalisability of The
30 Health Improvement Network (THIN) database: demographics, chronic disease
31 prevalence and mortality rates. *Inform Prim Care* 2011;251-255.

32
33
34
35
36
37 [24-29](#) National Centre for Social Research and University College London.
38 Department of Epidemiology and Public Health, *Health Survey for England, 2008*
39 [computer file]. *2nd Edition*. Colchester, Essex: UK Data Archive [distributor],
40 October 2010. SN: 6397

41
42
43
44
45
46 [30 Booth N What are the Read Codes? *Health Libr Rev* 1994;177-82](#)

1
2
3
4
5
6
7
8 [25-31](#) Noble M, Wright G, Dibben C et al Indices of Deprivation 2004. Report to
9 the Office of the Deputy Prime Minister. London: Neighbourhood Renewal Unit;
10 2004

11
12
13
14 (<http://webarchive.nationalarchives.gov.uk/20120919132719/http://www.communi->
15 [ties.gov.uk/documents/communities/pdf/131209.pdf](http://www.communities.gov.uk/documents/communities/pdf/131209.pdf)) (Accessed January 2014)
16
17

18
19
20 [32 Stata Corporation. Stata Statistical Software: Release 11. College Station, TX:](#)

21 [Stata Corporation; 2009](#)

22
23
24
25
26 [33 Royston P Multiple imputation of missing values: Update of ice. Stata Journal](#)
27 [2005;527-536.](#)
28
29

30
31
32 [26-34](#) Spratt M, Carpenter J, Sterne JAC et al Strategies for Multiple Imputation
33 in Longitudinal Studies *Am. J. Epidemiol.* 2010;478-487
34
35

36
37
38 [27-35](#) White IR, Royston P, Wood A Multiple imputation using chained equations:
39 issues and guidance for practice (tutorial). *Stat Med* 2011;377-399
40
41

42
43
44 [28-36](#) Graham JW Missing data analysis: Making it work in the real world *Annu*
45 *Rev Psychol* 2009;549-576
46
47

48
49 ~~[29 Stata Corporation. Stata Statistical Software: Release 11. College Station, TX:](#)~~

50 ~~[Stata Corporation; 2009](#)~~
51
52
53
54

1
2
3
4
5
6
7
8
9
10 ~~30 Royston P Multiple imputation of missing values: Update of ice. *Stata Journal*~~
11 ~~2005;5:27-536.~~

12
13
14
15
16 ~~31-37~~ Rubin DB *Multiple imputation for non-response in surveys*. New York: John
17 Wiley and Sons; 1987

18
19
20
21
22 ~~32 Szatkowski L, Lewis S, McNeill A et al Can data from primary care medical~~
23 ~~records be used to monitor national smoking prevalence? *J Epidemiol*~~
24 ~~*Community Health* 2011. doi:10.1136/jech.2010.120154~~

25
26
27
28
29 ~~33 Shipton D, Tappin DM, Vadiveloo T et al Reliability of self reported smoking~~
30 ~~status by pregnant women for estimating smoking prevalence: a retrospective,~~
31 ~~cross sectional study *Br Med J* 2009;b4347.~~

32
33
34
35
36
37 ~~34-38~~ The British Medical Association and NHS Employers Quality and
38 Outcomes Framework guidance for GMS contract 2011/12; 2011
39 ([http://www.nhsemployers.org/Aboutus/Publications/Documents/QOF_guidance_](http://www.nhsemployers.org/Aboutus/Publications/Documents/QOF_guidance_GMS_contract_2011_12.pdf)
40 [GMS_contract_2011_12.pdf](http://www.nhsemployers.org/Aboutus/Publications/Documents/QOF_guidance_GMS_contract_2011_12.pdf)) (Accessed January 2014)

41
42
43
44
45
46
47 ~~35-39~~ Campbell S, Reeves D, Kontopantelis E et al Quality of Primary Care in
48 England with the Introduction of Pay for Performance *N Engl J Med* 2007;181-
49 190

1
2
3
4
5
6
7
8
9
10 | [36-40](#) Wardle H, Mindell J Adult cigarette smoking. In Craig R, Shelton N (Ed.),
11 | Health Survey for England 2007. Volume 1. Healthy lifestyles: Knowledge,
12 | attitudes and behaviour. 2008;149-176 NHS Information Centre.
13
14
15
16
17

18 | [37-41](#) Doll R, Peto R, Boreham J et al Mortality in relation to smoking: 50 years'
19 | observations on male British doctors *Br Med J* 2004,
20 | doi:10.1136/bmj.38142.554479.AE
21
22
23
24
25

26 | [38-42](#) Kenfield SA, Wei EK, Rosner BA et al Burden of smoking on cause-
27 | specific mortality: application to the Nurses' Health Study *Tob Control* 2010;248-
28 | 254
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

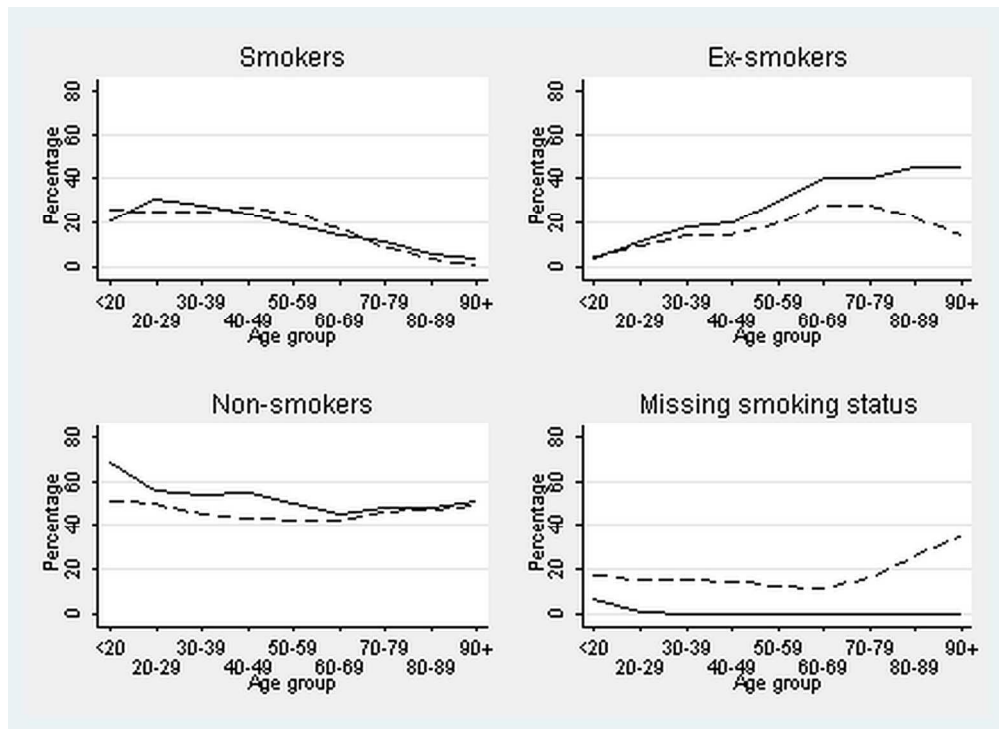


Figure 1: Smoking status percentages in THIN and the HSE 2008 by age group

Solid line is the Health Survey for England 2008, dashed line is The Health Improvement Network (THIN)

90x65mm (300 x 300 DPI)

ew only

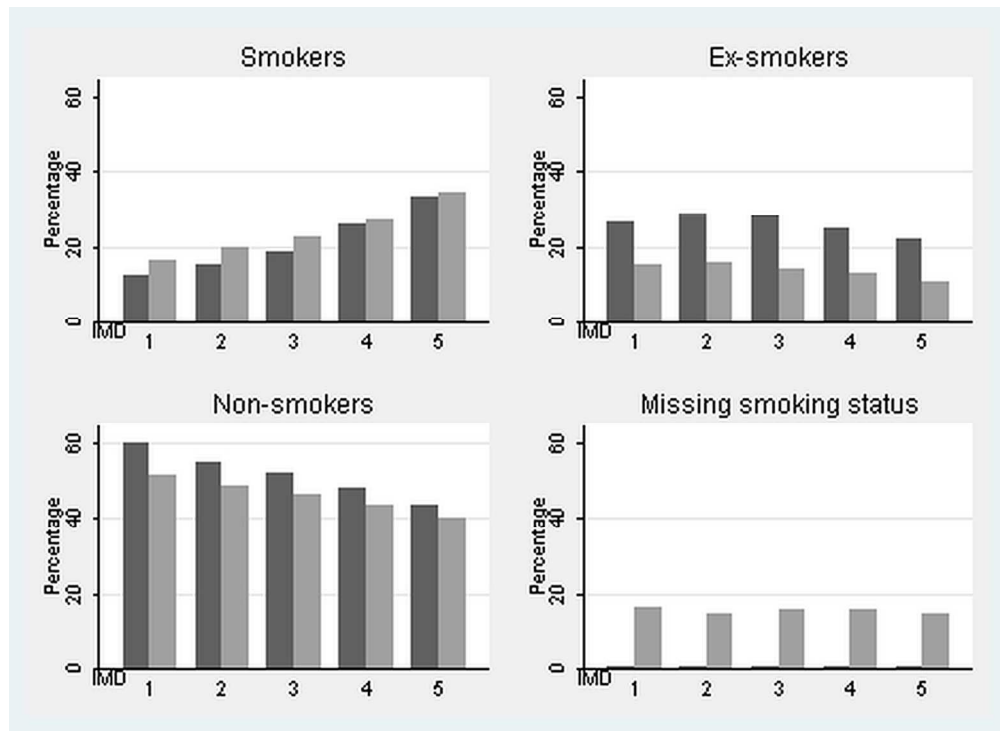


Figure 2: Smoking status percentages in THIN and the HSE 2008 by deprivation quintile

*IMD 1 is the least deprived and IMD 5 is the most deprived

Darker bars represent the HSE 2008, lighter bars represent THIN

Abbreviations: HSE Health Survey for England 2008, IMD Index of multiple deprivation, THIN The Health Improvement Network

90x65mm (300 x 300 DPI)

only

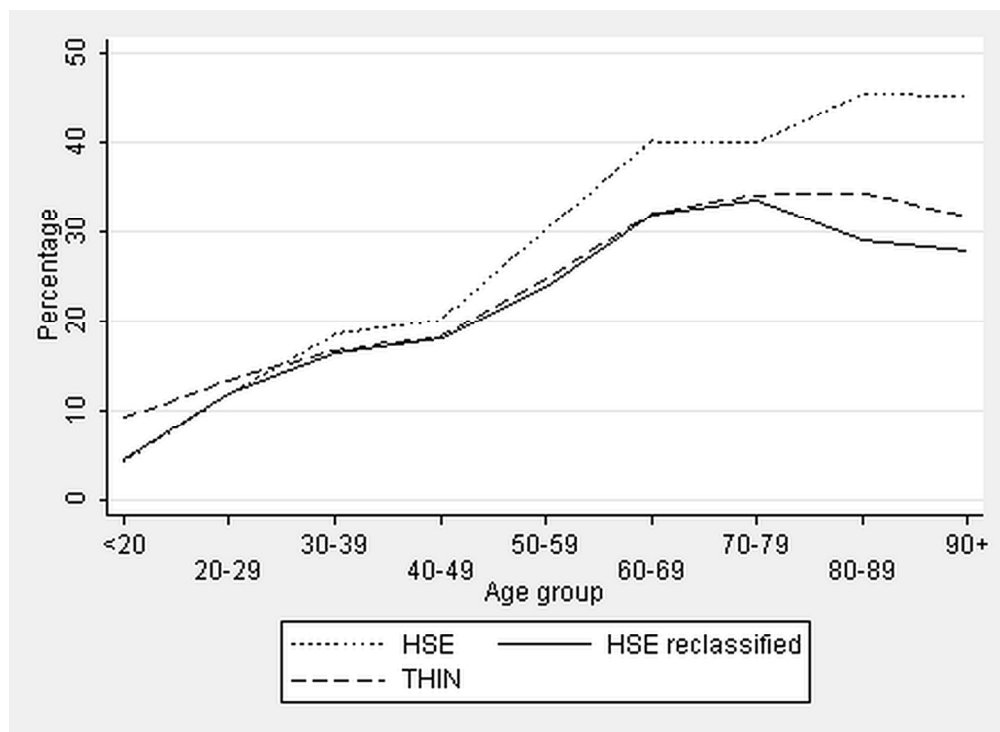


Figure 3: Age group specific percentages of ex-smokers in THIN (after MNAR imputation) and the HSE 2008 (before and after reclassifying ex-smokers in the HSE who quit before the age specified in Table 3 column 3 to be non-smokers)

Abbreviations: THIN The Health Improvement Network, HSE Health Survey for England 2008

90x65mm (300 x 300 DPI)

Appendix 1

For the Health Survey for England 2008 analysis, we used the derived variable cigsta3. This is a three category variable non-smokers, ex-smokers, current smokers. It is derived from a number of questions asked in the Health Survey for England as follows:

If the participant

- Reported never smoking in the constituent questions (“May I just check, have you ever smoked a cigarette, a cigar or a pipe?”, “Have you ever smoked cigarettes?” (the latter only asked to those who have smoked but does not smoke cigarettes nowadays)), they were coded as a never regular smoker on this variable.
- Using: “Did you smoke cigarettes regularly, that is at least one cigarette a day, or did you smoke them only occasionally?”, if participants answered “regularly”, they are coded as an ex-smoker. If they responded “occasionally” or “only tried once or twice”, they were categorised as a never regular smoker
- If participants answered “yes” to “Do you smoke cigarettes at all nowadays?” this was taken as “current smoker” in the variable cigsta3.
- If participants gave no answer/ refused to any of the constituent questions, this was carried forward to cigsta3
- If participants responded don’t know to any of the constituent questions, this was carried forward to cigsta3
- If participants answered not applicable to “Do you smoke cigarettes at all nowadays?” this was carried forward to cigsta3.

National Centre for Social Research and University College London. Department of Epidemiology and Public Health, *Health Survey for England, 2008* [computer file]. *2nd Edition*. Colchester, Essex: UK Data Archive [distributor], October 2010. SN: 6397

STROBE 2007 (v4) Statement—Checklist of items that should be included in reports of *cross-sectional studies*

Section/Topic	Item #	Recommendation	Reported on page #
Title and abstract	1	(a) Indicate the study's design with a commonly used term in the title or the abstract	1
		(b) Provide in the abstract an informative and balanced summary of what was done and what was found	3-4
Introduction			
Background/rationale	2	Explain the scientific background and rationale for the investigation being reported	5-7
Objectives	3	State specific objectives, including any prespecified hypotheses	7
Methods			
Study design	4	Present key elements of study design early in the paper	7-8
Setting	5	Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection	7-8
Participants	6	(a) Give the eligibility criteria, and the sources and methods of selection of participants	8
Variables	7	Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable	8-10
Data sources/ measurement	8*	For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group	8-10
Bias	9	Describe any efforts to address potential sources of bias	8
Study size	10	Explain how the study size was arrived at	NA
Quantitative variables	11	Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen and why	9-10
Statistical methods	12	(a) Describe all statistical methods, including those used to control for confounding	9-11
		(b) Describe any methods used to examine subgroups and interactions	9-11
		(c) Explain how missing data were addressed	9-11
		(d) If applicable, describe analytical methods taking account of sampling strategy	NA
		(e) Describe any sensitivity analyses	NA
Results			

Participants	13*	(a) Report numbers of individuals at each stage of study—eg numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analysed	11
		(b) Give reasons for non-participation at each stage	NA
		(c) Consider use of a flow diagram	NA
Descriptive data	14*	(a) Give characteristics of study participants (eg demographic, clinical, social) and information on exposures and potential confounders	11-12
		(b) Indicate number of participants with missing data for each variable of interest	12
Outcome data	15*	Report numbers of outcome events or summary measures	12
Main results	16	(a) Give unadjusted estimates and, if applicable, confounder-adjusted estimates and their precision (eg, 95% confidence interval). Make clear which confounders were adjusted for and why they were included	12-16
		(b) Report category boundaries when continuous variables were categorized	NA
		(c) If relevant, consider translating estimates of relative risk into absolute risk for a meaningful time period	NA
Other analyses	17	Report other analyses done—eg analyses of subgroups and interactions, and sensitivity analyses	NA
Discussion			
Key results	18	Summarise key results with reference to study objectives	16
Limitations	19	Discuss limitations of the study, taking into account sources of potential bias or imprecision. Discuss both direction and magnitude of any potential bias	17-19
Interpretation	20	Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence	17-18
Generalisability	21	Discuss the generalisability (external validity) of the study results	19
Other information			
Funding	22	Give the source of funding and the role of the funders for the present study and, if applicable, for the original study on which the present article is based	21

*Give information separately for cases and controls in case-control studies and, if applicable, for exposed and unexposed groups in cohort and cross-sectional studies.

Note: An Explanation and Elaboration article discusses each checklist item and gives methodological background and published examples of transparent reporting. The STROBE checklist is best used in conjunction with this article (freely available on the Web sites of PLoS Medicine at <http://www.plosmedicine.org/>, Annals of Internal Medicine at <http://www.annals.org/>, and Epidemiology at <http://www.epidem.com/>). Information on the STROBE Initiative is available at www.strobe-statement.org.