**Can different primary care databases produce comparable estimates of burden of disease:
Results of a study exploring venous leg ulceration.**

Emily S Petherick[1,2], Kate E Pickett[2,3], Nicky A Cullum[4]

1. Bradford Institute for Health Research, Bradford, UK
2. School of Health Studies, University of Bradford, Bradford, UK
3. Department of Health Sciences, University of York, York, UK
4. Hull York Medical School, University of York, York, UK
5. School of Nursing, Midwifery and Social Work, University of Manchester

Background: Primary care databases from the UK have been widely used to produce evidence on the epidemiology and health service usage of a wide range of conditions. To date there have been few evaluations of the comparability of estimates between different sources of these data.

Aim: To estimate the comparability of two widely used primary care databases, the Health Improvement Network Database (THIN) and the General Practice Research Database (GPRD) using venous leg ulceration as an exemplar condition.

Design of study: Cross prospective cohort comparison.

Setting: General Practice Research Database(GPRD) and the Health Improvement Network (THIN) databases using data from 1998 to 2006.

Method: A datasets was extracted from both databases containing all cases of persons aged 20 years or greater with a database diagnosis of venous leg ulceration recorded in the databases for the period 1998 to 2006. Annual rates of incidence and prevalence of venous leg ulceration were calculated within each database and standardised to the European standard population and compared using standardised rate ratios.

Results: Comparable estimates of venous leg ulcer incidence from the GPRD and THIN databases could be obtained using data from 2000 to 2006 and of prevalence using data from 2001 to 2006.

Conclusions: Recent data collected by these two databases is more likely to produce comparable results of the burden venous leg ulceration. These results require confirmation in other disease areas to enable researchers to have confidence in the comparability of findings from these two widely used primary care research resources.

In the UK most health care is provided within primary care, making primary care databases an ideal source for the studying health conditions in ambulatory populations. Two of the largest databases with long durations of follow up are the Health Improvement Network (THIN) and the Clinical Practice Research Datalink (CPRD), formerly known as the General Practice Research (GPRD) database. As few comparisons have been made between the two databases the comparability of disease estimates between these databases remains unclear (1, 2).

Leg ulcers cause a significant health burden with prevalence studies showing that approximately one per cent of people suffer from the condition, with higher prevalence in older people and women(3) and recurrence rates of up to 67 per cent (4). Most leg ulcers are due to underlying venous disease and cause significant pain and reduced quality of life and their management presents a significant cost burden with recent estimates suggesting costs of £400 million per annum to the National Health Service(5). If health care planning decisions are to be based on the findings of results of from database studies it is crucial to be aware of any inaccuracy that may be present in results, potentially due to differences in the age and gender structure within different database populations in addition to temporal differences.

There have been two previous studies using primary care databases to study the epidemiology of venous leg ulceration. Firstly, Margolis et al. (6) used the GPRD to determine incidence and prevalence of venous leg ulceration in the UK population aged 65 years and over from 1988 to 1996 confirming that large numbers of people sought treatment for leg ulceration in general practice. Furthermore, Margolis et al (6) validated venous leg ulcer coding in the GPRD and found it to have high sensitivity and specificity. More recently we undertook analyses in another UK primary care database, the THIN database, exploring the burden of venous leg ulceration as well as examining the implementation of guideline recommended care for leg ulcer patients. (7)

Estimates provided by these earlier crude estimates, produce results applicable to the populations in which they were studied but do not allow unbiased comparison as rates produced may mask differences that due to differences in age structures between populations. A strategy used to enable comparisons between populations and different temporal periods direct standardisation. Direct standardisation allows comparison of rates from multiple studies to be compared by choosing an appropriate reference standard to which the rates from different age strata of the two populations can be compared(8), thus eliminating the confounding effects caused by differences in age distributions between populations.

The problematic nature of conducting research using primary care databases has however been highlighted. In particular Muller noted in an editorial that '*A major criticism from peer reviewers of papers using EMR data is the potential for inaccuracies in diagnosis*'(9).

Although both the GPRD and the THIN databases have been used to conduct numerous burden of disease studies of many conditions including leg ulceration, none have attempted to compare estimates provided by them. The aim of this study was to determine the comparability of incidence and prevalence estimates of venous leg ulceration between two of the UK's largest general practice databases. Whilst this study will not allow assessment of the diagnostic accuracy of the diagnosis in these two databases it will enable exploration of the temporal consistency and comparability of diagnostic estimates from different primary care databases.

Methods

Case ascertainment

Cases were ascertained for the current study using the Read codes described by Margolis et al. (6) in his GPRD study of venous leg ulceration. For the GPRD analyses only, which historically we were further able to search using OXMIS codes described by Margolis et al. (6) which were used to code historically. The codes used to identify cases are shown in table 1 below.

Insert table 1 here

Inclusion criteria

All data meeting the 'up to standard' quality criteria of the GPRD or the acceptable mortality reporting standard (AMR) of the THIN database were used included if venous leg ulcers were recorded in the databases from January 1988 and December 2006 and came from patients aged 20 years or greater.

Additional inclusion criteria for the calculation of incidence

For the incident cohort, the inclusion criteria used by Margolis et al. (6) and Petherick et al (7) were applied as were the methods used to calculate the denominator of person years at risk. In brief this entailed cases only being included as incident if i) the initial diagnosis of leg ulceration was made at least six months after the commencement of the patient's database record and ii) there was no diagnosis of any other form of leg or foot ulcer recorded in the three months after the initial diagnosis.

Method for calculation of average annual incidence density

Annual incidence density over the study period was calculated using the formula and were presented per 100,000 person years.

$$\frac{\text{Number of new cases for each year between 1988 and 2006}}{\text{Number of person years at risk for each year between 1988 and 2006}}$$

Methods for the calculation of average annual period prevalence

Annual period prevalence was calculated using the formula shown below and presented per 100,000 persons at risk.

$$\frac{\text{Annual cases of for each year between 1988 and 2006}}{\text{Annual population at risk for each year between 1988 and 2006}}$$

Methods for the calculation of standardised incidence and prevalence rates.

Estimates of incidence and prevalence were then standardised to the European standard population(10) which is the same for both genders. Confidence intervals and standard errors for estimates of rates were calculated using methods described by Breslow & Day (8). The standardised rates calculated from both databases were compared by calculating a directly standardised rate ratio (SRR), defined as the ratio between the standardised rates in the GPRD divided by the standardised rates from the THIN database using methods described by Miettinen (11).

Results

Results of incidence

The records of patients that met the criteria as an incident case were extracted from the two primary care databases and examined.

GPRD results

The original data set supplied by the GPRD contained the records of 61,068 patients with a database diagnosis of venous leg ulceration and of these, 37,575 or 61.5% met the inclusion criteria as an incident case and were considered for further analysis.

THIN results

The dataset supplied by the THIN database contained the records of 22,788 patients with a database diagnosis of venous leg ulceration and of these 20,261 or 88.9% met the inclusion criteria as an incident case and were included for further analysis.

Summary characteristics of incident cohort

The baseline characteristics of incident leg ulcer patients identified in the two databases are shown below in table 2. The results are presented stratified by the database where the events had been recorded.

Insert table 2 here

The investigation of incident venous leg ulceration was conducted over the same time period in both databases, from 1988 to 2006. The demographic characteristics of patients showed little variation between the two databases. The crude estimate of the incidence density rate of venous leg ulceration over the study period was 122 per 100,000 person years (95% C.I. 120.7-123.2) in the GPRD and 81 per 100,000 person years (95% C.I. 79.9-82.2) in the THIN database. Crude annual incidence density rates of venous leg ulceration obtained from the GPRD and THIN databases were compared and shown below in figure 1. Differences between the crude rates of the incidence density of venous leg ulceration that were evident early in the study period, were shown to diminish considerably from the year 2000 onwards.

Insert figure 1 here

To further test whether rates from these two databases were comparable, the crude incidence density rates of venous leg ulceration obtained were standardised. These results are shown below in figure 2.

Insert figure 2 here.

Standardised incidence rates were shown to follow a very similar temporal pattern to the crude rates although the actual estimates produced were lower. Once again, rates from the GPRD were shown to peak in 1990 whilst rates in the THIN database remained considerably lower until 2000.

Lastly standardised rate ratios (SRR) were calculated to statistically compare the standardised venous leg ulcer incidence density rates from the GPRD and THIN databases. These results indicated that the estimates of venous leg ulcer incidence obtained from the two databases were not statistically significantly different from the year 2000 onwards (see table S1 in supplemental material).

Results of prevalence
A summary of the results of prevalence from each of the databases is provided below.

GPRD prevalence results
The original data set supplied by the GPRD contained the records of 61,068 patients with a database diagnosis of venous leg ulceration. Of these patients, 47,760 or 78.2% met the inclusion criteria as a prevalent case during the study period of January 1988 to December 2006 and were considered for further analyses.

THIN prevalence results
The dataset supplied by the THIN database contained the records of over 22,788 patients with a database diagnosis of any form of leg ulceration. Of these patients, 20,619or 90.4% met the inclusion criteria as a prevalent case during the study period and were included for further analyses.

Summary characteristics of the prevalent cohort
The baseline characteristics of the prevalent leg ulcer cohort identified in both databases are shown below in table 3. The results are presented stratified by the database diagnosis of leg ulceration and database location.

Insert table 3 here

The characteristics of the prevalent cohort by database leg ulcer diagnosis were quite consistent in both databases. In common with the earlier results for incidence, greater numbers of women had leg ulcers compared to men. The mean and median ages of prevalent cases were higher than those observed earlier for incident cases as would be expected with a chronic recurrent condition such as leg ulceration.

Crude annual prevalence rates of venous leg ulceration were calculated and the results between the databases compared. The result of this comparison is shown below in figure 3.

Insert figure 3 here

During the years 1988 through to 1999, crude annual prevalence rates of venous leg ulceration from the GPRD were higher than the crude annual rates from the THIN database. From 2000 until the end of the study period in 2006, crude annual rates between the two databases showed little variation. Rates over the entire period ranged from 82.8 per 100,000 person years in the THIN database (95% C.I. 81.7-83.9) to 140.7 per 100,000 person years in the GPRD (95% C.I 139.5-142.0).

Insert figure 4 here

Crude rates from both databases were standardised and the results of this analysis are shown above in figure 4. Standardised rates of venous leg ulcer prevalence showed similar temporal patterns to the earlier crude results although the estimated rates were approximately half of the crude results. The results demonstrated that there were large differences in rates between the databases over the period 1988 to 1999. Standardised rates from 2000 onwards narrowed the results between the two databases further than the crude results. Results within the time period 2000 to 2006 were never than more than 20 per 100,000 persons different between the databases. By 2006 the difference between the results from both databases had once again narrowed to 2 per 100,000 persons.

Standardised rate ratios were calculated to examine any potential differences in the prevalence rates between the two primary care databases. The results, shown supplementary table 2 indicated that comparable results of venous leg ulcer prevalence between the two databases could be obtained from the year 2001 onwards.

Discussion

Two of the largest UK primary care databases were searched to identify venous leg ulcer patients with a database record of incident or prevalent ulceration, consulting during the study period of January 1988 to December 2006. This search located over 56,000 incident leg ulcer patients and over 67,000 prevalent patients.

These results indicated that comparable rates of annual incidence rates venous leg ulcers could be obtained from the GPRD and THIN databases after year 2000 only and for prevalence between 2001 and 2006. These results indicated that there were statistically significant results between the databases for the majority of the eighteen year study period investigated, although these differences diminished from the year 2000 onwards. The exploration of venous leg ulcer disease burden trend over time should therefore

be limited to these recent data to exclude the possibility of bias caused by extrinsic differences in estimates of venous leg ulcer burden between the two databases.

There are several reasons that may explain differences in leg ulcer burden of disease estimates between the GPRD and THIN databases prior to 2000. Irrespective of the primary care database that they contribute data to, new practices are more likely provide incomplete data as they learn to use new computer systems and achieve new quality standards of clinical data reporting. During the study period investigated, more practices have joined the THIN database, including half of those that also contribute to the GPRD. In contrast more practices have stopped contributing to the GPRD and <u>was seen by the amount of leg ulcer cases being excluded from prevalence estimates more historical data was included</u>. By the year 2000, more of those practices that joined the THIN database had contributed data electronically for several years and had met the acceptable mortality reporting standard required by the database. These factors are the likely cause of the comparable estimates that have been observed in this study. A further explanation for the observed results may be differences in both the coding, software and recording of data used between the two databases. Data on all historical clinical and diagnostic events from the THIN database has been converted into Read codes whereas in contrast the GPRD has kept the combination of Read and historical OXMIS coded events, which in the current study included events up until 1999, although only Read codes have been used from this period onwards (Penn, 2010).

Strengths and weaknesses of the current study

This study was undertaken using two of the largest general practice databases available in the UK containing longitudinal data to examine the comparability of venous leg ulcer disease burden trends over time. The current study used a previously validated case ascertainment strategy to identify patients with venous leg ulceration. This was found to be sufficiently sensitive and specific when identifying patients in the GPRD (6). Although this study was limited as no leg ulcer validation studies have been undertaken in the THIN database, it is known that approximately half of all practices that provide data to the GPRD also contribute to the THIN database (12). In both databases there is the possibility that there may be misclassification of the database venous leg ulcer diagnosis although it is unlikely that there was non-differential misclassification of leg ulcer diagnoses between practitioners contributing to the database that would have produced the observed results.

There are several methodological advantages to using primary care databases to derive burden of disease estimates for venous leg ulceration. First, this approach avoids any non-response bias that was evident in many earlier studies where case ascertainment was dependent upon surveying health professionals to identify leg ulcer patients (3). In some studies fewer than 50% of health professionals responded to requests for details of their leg ulcer patient population (13). Second, the methods we used ensured that all results were based on prospectively collected clinical data that had met stringent quality standards.

Third, the strength of this approach means that data were not subject to recall error or selection bias from either patients or practitioners. Patients may find it difficult to remember when they were first diagnosed with leg ulceration or for how long they have had the condition, particularly due to the chronic recurrent nature of the condition. Recall error was therefore eliminated in the current study as the data obtained comes from prospectively collected primary care medical records. Finally selection bias is also excluded as all patients records can be accessed so there is no chance of any patient's records being systematically excluded. A limitation of the current study is that due to the way practitioners code clinical events in primary care, prevalence may be underestimated, as it has been noted that many clinical events may only be recorded on their first occurrence and not for subsequent episodes if there is no change in clinical management(14). There is the possibility that

Strengths and weaknesses in relation to other work

The crude annual prevalence rates of venous leg ulceration calculated in this study are broadly similar to previous estimates from studies conducted in the general adult population with crude prevalence estimates reported by Graham et al. of 1.1% of the population with open ulcers and 0.9 per 100 people obtained in this study (3). The results from this current study show that the time period over which Margolis et al. (6) produced their results corresponded with the highest rates observed over the entire study period of 1988 and 2006 of the current study and the greatest differences between the GPRD and THIN databases.

There is some evidence from other disease areas, including cancer, indicating that incidence rates from the THIN database may be higher than those observed from the CPRD (formerly known as GPRD). Afonso et al (2013) compared both crude and age standardised rates of all forms of cancer during the years 2001 to 2009 and found that rates were consistently higher in the THIN compared to the CPRD(2). As other comparative studies were not found it remains unclear whether these results apply to other disease areas but do show a consistent pattern with our results. A systematic review undertaken to specifically examine the validity of coding in the GPRD found inconsistent replication consultation or prevalence rates compared with other primary care databases in the UK and in the US, in studies considering a wide range of different disease areas(15). More recently studies have used the THIN database to compare rates of smoking during pregnancy with other survey data and found good concordance with published estimates of congenital anomalies, although rates of smoking only showed agreement with other data after 2004(16, 17).

The aetiological classification of venous leg ulceration used in this study came from the Read code or in the case of the GPRD Read and OXMIs codes, assigned by the treating health professional which may be unreliable. The proportion of patients diagnosed as having venous leg ulceration in this study may therefore be overestimated as an earlier study found that venous leg ulcer patients treated within primary

care may not routinely be provided with recommended Doppler ultrasound assessment which would aid diagnosing of the underlying pathology of leg ulceration(7). Despite these limitations the case ascertainment strategy used for venous leg ulceration has previously been validated and found to be reliable (6).

This study accessed patients' retrospective medical records meaning that there was no selection, systematic reporting bias or recall error that may be present in studies that relied on health care professionals to provide details of leg ulcer patients. We did not however have access to full patient populations of both databases, so it is unclear if the results observed are down to differences in population structures of the contributing patient populations over time, i.e. the GPRD had more elderly participants or different sex distribution at certain time points. If this had however been the case we would have expected that standardisation would have diminished some of this differences, which was not what we observed.

In conclusion this study has found that consistent and comparable estimates of venous leg ulcer burden can be obtained from both the GPRD and THIN databases from the year 2000 onwards for estimates of incidence and 2001 onwards for estimates of prevalence. Data from these time periods can be used to gain comparable data of both the epidemiology and management of leg ulceration. The results from this study further highlight the caution that must be applied when making comparisons of coding of clinical events between primary care databases *prima facie*. Primary care database studies are a powerful resource with which to derive timely and comprehensive intelligence about the health burden and utilisation for conditions that are treated within the primary care setting. Replication of these findings is required to examine the generalizability to other health outcomes.

References

1.   Lewis JD, Schinnar R, Bilker WB, Wang X, Strom BL. Validation studies of the health improvement network (THIN) database for pharmacoepidemiology research. *Pharmacoepidemiol Drug Saf.* 2007;**16**(4):393-401
2.   Afonso AS, De Groot MCH, Van De Ham R, De Bruin ML, Hesse U, et al. Comparison of Six Electronic Healthcare Databases in Europe Using Standardized Protocols: A Descriptive Study on the

Incidence of Cancer. International conference on Pharmacoepidemiology & Therapeutic Risk Management; 7/10/2013; Montreal, Canada: Pharmacoepidemiology & Drug Safety; 2013. p. 285.

3.   Graham ID, Harrison MB, Nelson EA, Lorimer K, Fisher A. Prevalence of lower-limb ulceration: a systematic review of prevalence studies. *Adv Skin Wound Care*. 2003;**16**(6):305-16

4.   Callam MJ, Harper DR, Dale JJ, Ruckley CV. Chronic ulcer of the leg: clinical history. *BMJ*. 1987;**294**(6584):1389-91

5.   Simon DA, Freak L, Kinsella A, Walsh J, Lane C, et al. Community leg ulcer clinics: a comparative study in two health authorities. *BMJ*. 1996;**312**(7047):1648-51

6.   Margolis DJ, Bilker W, Santanna J, Baumgarten M. Venous leg ulcer: incidence and prevalence in the elderly. *Journal of the Amercian Academy of Dermatology*. 2002;**46**(3):381-6

7.   Petherick ES, Cullum NA, Pickett KE. Investigation of the effect of deprivation on the burden and management of venous leg ulcers: a cohort study using the THIN database. *PLoS One*. 2013;**8**(3):e58948

8.   Breslow NE, Day NE. Statistical Methods in Cancer Research: Vol. II - The Design and Analysis of Cohort Studies. Lyon: International Agency for Cancer Research; 1987.

9.   Muller S. Electronic medical records: the way forward for primary care research? *Family practice*. 2014

10.   Waterhouse J, Muir C, Correa P, Powel J, editors. Cancer Incidence in Five Continents. Volume III. IARC Scientific Publication no.15. Lyon: World Health Organisation; 1976.

11.   Miettinen OS. Standardization of risk ratios. *Am J Epidemiol*. 1972;**96**(6):383-8

12.   The Health Improvement Network. THIN data from EPIC, A guide for researchers. 2007.

13.   Hickie S, Ross S, Bond C. A survey of the management of leg ulcers in primary care settings in Scotland. *J Clin Nurs*. 1998;**7**(1):45-50

14.   Hansell A, Hollowell J, Nichols T, McNiece R, Strachan D. Use of the General Practice Research Database (GPRD) for respiratory epidemiology: a comparison with the 4th Morbidity Survey in General Practice (MSGP4). *Thorax*. 1999;**54**(5):413-9

15.   Khan NF, Harrison SE, Rose PW. Validity of diagnostic coding within the General Practice Research Database: a systematic review. *The British journal of general practice : the journal of the Royal College of General Practitioners*. 2010;**60**(572):e128-36

16.   Sokal R, Fleming KM, Tata LJ. Potential of general practice data for congenital anomaly research: Comparison with registry data in the United Kingdom. *Birth defects research Part A, Clinical and molecular teratology*. 2013;**97**(8):546-53

17.   Dhalwani NN, Tata LJ, Coleman T, Fiaschi L, Szatkowski L. A comparison of UK primary care data with other national data sources for monitoring the prevalence of smoking during pregnancy. *Journal of public health*. 2014

Tables

Table 1 Comparison of codes used to identify the venous leg ulcer cohort in both the GPRD and THIN databases

GPRD Medical Code	Read / OXMIS Code	Read / OXMIS Term	Used in the THIN database	GPRD (%)	THIN (%)
216082	G830.00	Varicose veins of the leg with ulcer	Yes		
345418	G832.00	Varicose veins of the leg with ulcer and eczema	Yes		
339862	G837.00	Venous ulcer of leg	Yes		
339887	14F5.00	H/O: venous leg ulcer	No, OXMIS code		
219441	K914 RR	EXCISION VARICOSE ULCER	No, OXMIS code		
271667	M271500	Venous ulcer of leg	Yes		
303889	4540	VARICOSE ULCER LEG	No, OXMIS code		
303890	4540N	VARICOSE ULCER	No, OXMIS code		
303892	4540NE	VENOUS ULCER	No, OXMIS code		
280021	G832.00	Varicose veins of the leg with ulcer and eczema	Yes		
303891	4540NA	ULCER STASIS VARICOSE	No, OXMIS code		
289131	G835.00	Infected varicose ulcer	Yes		
256627	4540A	ULCER VARICOSE INFECTED (LEG)	No, OXMIS code		
262397	M271.00	Non-pressure ulcer lower limb	Yes		
256936	707 GL	ULCER LOWER LEG	No, OXMIS code		
235019	M271.13	Leg ulcer NOS	Yes		
304723	707 G	ULCER LEG	No, OXMIS code		
304724	707 GA	ULCER ANKLE	No, OXMIS code		
304718	707 AC	ULCER SKIN	No, OXMIS code		
256937	707L	ULCER GRAVITATIONAL CHRONIC	No, OXMIS code		
256935	707 AL	ULCER LOWER EXTREMITY	No, OXMIS code		
304719	707 A	ULCER SKIN CHRONIC	No, OXMIS code		

Table 2 Baseline characteristics of the incident cohort

	GPRD	THIN
Ulcer type, N(%)		
Venous	37575 (98.2)	19378 (96.0)
Patient characteristics		
Female N(%)	24830 (65.0)	12870 (63.5)
Mean age (SD)	73.2 (14.4)	73.2 (14.1)
Median age, range	76 (18-109)	76 (18-109)

Table 3 Baseline characteristics of the prevalent cohort

	GPRD	THIN
Ulcer type, N(%)		
Venous	47760 (97.7)	20619 (94.3)
Patient characteristics		
Female N(%)	31767 (65.0)	13336 (64.6)
Mean age (SD)	74.0 (14.3)	73.8 (14.2)
Median, range	77 (18-109)	77 (18-109)