Family Practice 2012; **29**:315–331
doi:10.1093/fampra/cmr100
Advance Access published on 3 February 2012

An international comparative family medicine study of the Transition Project data from the Netherlands, Malta, Japan and Serbia. An analysis of diagnostic odds ratios aggregated across age bands, years of observation and individual practices

Jean K Soler[a,b,*], Inge Okkes[b,c], Sibo Oskam[c], Kees van Boven[d], Predrag Zivotic[e], Milan Jevtic[f], Frank Dobbs[a,b] and Henk Lamberts[c] for the Transition Project

[a]Faculty of Life and Health Sciences, University of Ulster, Coleraine, UK, [b]Mediterranean Institute of Primary Care, Attard, Malta, [c]Formerly of the Department of General Practice, Academic Medical Centre, University of Amsterdam, Amsterdam, The Netherlands, [d]Department of Primary and Community Care, Nijmegen Medical Centre, Radboud University, Nijmegen, the Netherlands, [e]BONEX inzenjering, Gandijeva Str. 148a, Belgrade 11070 and [f]Saga System Integration, 11070 Belgrade, Serbia.
*Correspondence to Jean K Soler, Mediterranean Institute of Primary Care, 19, Triq ir-Rand, Attard ATD1300, Malta;
E-mail: jksoler@synapse.net.mt

Received 9 May 2011; Revised 28 September 2011; Accepted 1 October 2011.

Introduction. This is a study of the process of diagnosis in family medicine (FM) in four practice populations from the Netherlands, Malta, Serbia and Japan. Diagnostic odds ratios (ORs) for common reasons for encounter (RfEs) and episode titles are used to study the process of diagnosis in international FM and to test the assumption that data can be aggregated across different age bands, practices and years of observation.

Methodology. Participating family doctors (FDs) recorded details of all their patient contacts in an episode of care (EoC) structure using the International Classification of Primary Care (ICPC). RfEs presented by the patient and the diagnostic labels (EoC titles) recorded for each encounter were classified with ICPC. The relationships between RfEs and episode titles were expressed as ORs using Bayesian probability analysis to calculate the posterior (post-test) odds of an episode title given an RfE, at the start of a new EoC.

Results. The distributions of diagnostic ORs from the four population databases are tabled across age groups, years of observation and practices.

Conclusions. There is a lot of congruence in diagnostic process and concepts between populations, across age groups, years of observation and FD practices, despite differences in the strength of such diagnostic associations. There is particularly little variability of diagnostic ORs across years of observation and between individual FD practices. Given our findings, it makes sense to aggregate diagnostic data from different FD practices and years of observation. Our findings support the existence of common core diagnostic concepts in international FM.

Keywords. Diagnosis, electronic medical records, electronic patient records, episode of care, family medicine, general practice, ICPC, international, International Classification of Primary Care, Japan, Malta, reason for encounter, Serbia, the Netherlands, Transition Project.

Introduction

The first two articles in this series studied differences and similarities in the distributions of utilization, reasons for encounter (RfEs) and episode titles and analysed diagnostic relationships expressed as odds ratios (ORs), in patient populations in three countries using data from the Transition Project.[1,2] These data were used to study the content of the discipline of family medicine (FM, synonymous with general practice), in an international comparison. These data evidenced important similarities in diagnostic relationships between the populations studied. The data in these first two papers were analysed at a population level, aggregating data from different practices, across age-sex groups and over a period of observation spanning

from 1 to 11 years. However, we did not test the assumption that data can be so aggregated.[1,2]

This paper now tests whether these 'diagnostic ORs' exhibit similar qualities, including differences and similarities when analysed across age groups, years of observation and practices, rather than aggregated between them. How robust are the conclusions we have made in the previous papers and do they hold with this different analytic approach?[1,2]

Most diagnostic studies pool data from different practices and from patients in different age groups who consult in the observation period. The published literature supports this, suggesting that the variation of many observed rates between practices in one year is less than between years for one practice, but this variation impacts differently on different observations (such as different body systems, types of interventions, prescriptions).[3] Boerma[4] reports that the variation between family practice service profiles is better explained by differences between health care systems than variability between family doctors (FDs, synonymous with GPs). Although possibly small, the effect of inter-doctor variation in diagnostic associations has not been widely studied in FM.

Similarities and differences in the content of family practice data from different populations reflect numerous effects, including that of inter-doctor variation. What is the effect of inter-doctor variation on diagnostic associations, and how may this affect the conclusions reached in our previous study of such data?[2]

A number of longitudinal electronic medical record (EMR) datasets collected from the daily practice of FDs are available to researchers, including databases collected from the Netherlands and the UK. Most of these databases preferentially collect information on the diagnostic label and are encounter based. Very few systematically collect data on the patient's RfE (defined in Methodology section) and structure data in the form of *episodes of care* (EoCs; defined in Methodology section). Such data elements greatly enhance the utility for diagnostic research, as we have demonstrated previously.[1,2,5] In previous publications, we used the Transition Project data from Malta, the Netherlands and Serbia, also to study diagnostic associations. Data from the Japanese arm of the Transition Project are included in this study, even though they are not concurrent, to broaden the scope of the comparisons to analyses of a smaller dataset.

This paper will focus specifically on the effects of patient age, the observation (time) window and individual practice on diagnostic associations expressed as ORs. Do the conclusions which we have previously made with respect to an international core diagnostic process in the discipline of FM still stand when the data are analysed in this way? Can a diagnostic relationship summarized by a likelihood ratio or OR in one population be used to support a diagnostic decision in another? The study will thus inform the methodology of aggregating FM diagnostic data and the utility of such analyses.

Research questions

- What are the quantitative relationships between common RfEs and common diagnoses (episode titles) within EoCs in routine family practice in practice populations from Malta, the Netherlands, Japan and Serbia, across age groups, periods of observation and practices?
- What are the generic similarities and differences in the relationships between common RfEs and common diagnoses (episode titles) in these practice populations?
- Do these similarities in the relationships between common RfEs and common diagnoses in this subanalysis support the existence of an international core process of diagnosis in the domain of FM?

Methodology

Data and setting

The freely available EMR TransHis[6] and the *International Classification of Primary Care* (ICPC; ICPC-1 in the Netherlands and Japan and ICPC-2 in Malta and Serbia)[7] were used to collect data. FDs participating in the Transition Project recorded details of all their patient contacts in a defined time period [158 370 patient-years during 11 years in the Netherlands (1995–2005), 43 577 patient-years during 5 years in Malta (2001–05), 17 042 patient-years during 3 years (1996–98) in Japan, 72 673 patient-years during 1 year in Serbia (2003)] in an EoC data structure. The populations in the Netherlands, Serbia and Japan represent registered patient populations (only those >15 years old in Serbia), while the population in Malta represents patients consulting over a 5-year period.[1,2] Some conditions are not usually seen by FDs in Serbia and Japan, as there is a requirement to refer cases directly to a specialist (e.g. ill children in Serbia and gynaecological problems in Japan).

Data elements

An EoC is defined as a health problem from its first presentation by the patient to the FD, until the completion of the last encounter for it. It encompasses all contact elements related to that health problem. Its name (i.e. the diagnostic label of the EoC) may be modified over time and is referred to as the 'episode title'.[7,8]

The RfE is defined as an agreed statement of the reason(s) why a person enters the health care system, representing the demand for care by that person. The RfE should be recognized by the patient as an acceptable description of the demand for care.[7,8] Doctors recording data for the Transition project were trained to record RfEs with ICPC according to the definitions above,

reflecting the patient's symptoms and requests as they expressed them. Symptoms elicited during history taking (i.e. the history of the presenting complaint) were recorded in a separate cell in the EMR Transhis and were not used for the analyses in this study.

ICPC has a biaxial structure, with 17 chapters on one axis and 7 components on the other.

Chapters are based on body systems, with an additional chapter for psychological problems and one for social problems.[7] Each chapter is identified by a single alphabetic code, which is the first character of all rubrics belonging to that chapter. Each chapter is divided into seven components, identified by a range of two digit numeric codes. Component 1 codes symptoms and complaints, while Component 7 codes diseases. An RfE can be either a symptom (Component 1) or a disease (Component 7) when a patient presents with an RfE such as 'doctor, I have migraine'. Conversely, an EoC may have a disease label diagnostic title, or it may be labelled with a Component 1 'symptom' diagnosis, such as when the FD cannot be more precise than label an EoC with the title 'shortness of breath'. Components 2–6 deal with interventions and can be used to code an RfE, which is presented as a request for an intervention.[7]

Analysis

The relationships between RfEs and diagnoses (episode titles) were studied using Bayesian probabilistic methods. According to Bayes' Theorem, the post-test (posterior) odds of an event (i.e. a diagnosis being made) is equivalent to the pre-test odds multiplied by the 'likelihood ratio'. The ORs presented in the tables are derived from these likelihood ratios and were calculated in a similar way to the method reported by Okkes *et al.*,[8,9] representing the odds of disease against no disease over odds of RfE present against absent, the RfE itself acting as a test. We modified the method slightly so as to calculate odds within an EoC rather than patient years of observation.[2] This has the advantage of estimating probabilities for a new problem at the beginning of an EoC.

The likelihood ratio is a mathematical expression of the extent to which a symptom increases the probability of a diagnosis. The (positive) likelihood ratio for the existence of the symptom is the odds that it will exist in a patient with the disease, in contrast to a patient without the disease. The (negative) likelihood ratio for absence of the symptom is the odds that a test will be negative in a patient with the disease, contrasted with a patient without the disease. The diagnostic ORs presented here are numerically equivalent to the positive likelihood ratio (LR+) divided by the negative likelihood ratio (LR–).

In this paper, the diagnostic ORs were calculated for different age groups, different practices and different years of observation, to study variation across observation frames. The case of finding small degrees of variability would support the aggregation of data, such as we have done in previous studies in this series.[2]

It would be possible to analyse such relationships between all possible combinations of episode titles and RfEs. The analysis was limited to selected examples for practical reasons. The examples chosen were two episode titles from the mental health chapter, namely 'depressive disorder' (P76) and 'anxiety disorder' (P74), and examples from the most prevalent ICPC chapter (R, respiratory). In the first case, the choice was made due to the fact that the diagnostic process in this area is often challenging and is based on symptoms rather than clinical signs and tests. The selection from chapter R was made to allow frequent observations with more data, namely the episode titles 'asthma' (ICPC rubric R96), 'acute tonsillitis' (R76) and the RfE 'wheezing' (R03).

Identification of diagnostic associations

In each case outlined above, the data from each Transition Project database were analysed to identify RfEs which could potentially have a significant association with the relevant episode title (in the first example, all symptoms (RfEs) which could make a contribution to making or excluding a diagnosis of depressive disorder at the start of a new EoC). For the RfE wheezing (R03), episode titles which could potentially have a significant diagnostic association were selected. This was done by calculating the standard error (SE) of observation of the rate for each RfE (expressed as a rate per 1000 observations) presenting for that episode title (or vice versa for wheezing) and discarding as unlikely to be significant all those associations where that SE was larger than half the size of the observation itself. This is numerically equivalent to the statistical significance limit for an OR, defined below as being at least as large as its confidence interval (CI).[2,3,8,9] All these potentially significant associations were further analysed in all four population databases, in all age groups, practice by practice (where data were available) and in each individual year of observation, by subsequently calculating the respective diagnostic OR for that association. If that OR was both clinically and statistically significant (see below) than it was highlighted in the table (bold type) as contributing to making the diagnosis for that episode title. For the RfE wheezing, associated episode titles were analysed in an analogous fashion.

Tables of diagnostic associations were thus drawn up for different populations, age groups, years of observation and individual practices, with their respective CIs. CIs were included to express our confidence limits in generalizing these observed diagnostic ORs to a larger 'population' of diagnostic decisions. Data for individual practices were only available from the Maltese and Dutch databases.

Clinical and statistical significance

The minimum level of 'clinical significance' for a diagnostic OR was arbitrarily taken as that which represents a standardized difference of at least 0.10 (10% of the variability is so explained). This is equivalent to a relative risk of ⩾2.0.[8–10] Since the OR tends to overestimate the relative risk, an arbitrary cut-off level of ⩾3 (rounded from ⩾2.45) for the OR of a positive association and ⩽0.3 (rounded from ⩽0.34) for the OR of a negative association were taken as thresholds for clinical significance. ORs which are outside these limits were still included in the tables but were not highlighted in bold type. Cells with very small numbers were ignored in the analysis.

Furthermore, ORs which are not at least as large as their CI were arbitrarily ignored as unreliable.[8,9] The strict criteria adjust for the increased chance of describing spurious associations due to the large numbers of statistical tests and for the effect of clustering of data on estimates of variance.[11]

Age groups

Age groups were taken as 0–14, 15–44, 45–64 and 65+. These four age bands were selected since narrower age bands (5-year or 10-year) would have resulted in wider CIs due to smaller numbers of observations.

Ethical approval

The study did not involve the collection of new data. Ethical approval was applied for locally, when appropriate, for individual studies based on these data in the Netherlands, Serbia Malta and Japan.

Results

We would suggest that a printed copy of all ICPC rubrics and short text labels might be useful while reading the Results and Discussion sections below. Such two-page documents are freely available in many languages from the Wonca website (http://www.globalfamilydoctor.com/wicc/pagers.html).

EoC 'depression' and its associated RfEs

Populations. Eight RfEs (Table 1) were found to have at least one clinically and statistically significant diagnostic OR for the episode title 'depression' (P76) across the four populations in the Netherlands (Nl), Malta (Mt), Serbia (Sb) and Japan (Jp): 'feeling depressed' (P03), 'feeling anxious' (P01), 'weakness/tiredness' (A04), 'sleep disturbance' (P06), 'headache' (N01), depression (P76) (i.e. a patient presenting with the RfE 'doctor, I have depression'), 'irritability' (P04) and 'acute stress' (P02). The significant ORs (highlighted in bold type) are in the same direction from unity, but vary in size.

Age groups. Table 2 gives the diagnostic ORs for the above associations, in different age groups. Significant ORs were observed mainly in the 15–64 age groups in the Dutch population and the 15–44 age group in the Maltese.

Years of observation. Table 3 gives the ORs in 1 year of observation, year by year. The ORs exhibit high degrees of congruency across years of observation.

Practices. Data for ORs across different FD practices were only available for the Netherlands and Malta (Table 4). Significant ORs were all in the same direction, as was the case for other observed ORs. However, some ORs had wider CIs than others. Some associations were only significant, or indeed observed, in individual FD practices, such as that for the RfE feeling depressed (P03) in Malta.

In-depth analysis. Details of the calculations behind the ORs, likelihood ratios as well as the sensitivity, specificity and a cross tabulation of the actual data for the RfE feeling depressed (P03) at the beginning of a new EOC of depression (P76) in the populations from the Netherlands (Nl), Malta (Mt), Serbia (Sb) and Japan (Jp) are given in Table 5. The positive and negative likelihood ratios are all greater than, or less than, unity respectively, with CIs which exclude unity.

TABLE 1 *RfEs with a significant association with the episode title 'depression' (P76), at the start of a new EoC (first encounter in a new episode)—populations*

RfE rubric	RfE label	Nl	Mt	Sb	Jp
P03	Feeling depressed	**152.37 (134.25–172.93)**	**348.00 (261.77–462.65)**	118.32 (50.36–278.03)	5175.00 (580.68–46 119.54)
P01	Feeling anxious/nervous/tense	**14.85 (12.43–17.74)**	**44.56 (33.24–59.73)**	10.39 (6.24–17.30)	–
A04	General weakness/tiredness	**4.81 (4.08–5.67)**	**5.93 (4.02–8.74)**	0.83 (0.12–5.96)	10.60 (4.00–28.11)
P06	Disturbances of sleep/insomnia	**10.10 (8.06–12.66)**	44.70 (26.82–74.51)	–	16.62 (4.96–55.71)
N01	Headache (excl N02 N89 R09)	0.55 (0.33–0.92)	**2.54 (1.57–4.10)**	–	3.30 (0.99–10.97)
P76	Depressive disorder	**1139.05 (883.43–1468.63)**	190.16 (11.87–3047.61)	–	–
P04	Feeling/behaving irritable	**24.43 (16.10–37.06)**	42.40 (9.12–197.08)	–	–
P02	Acute stress/trans/situat distur	**6.89 (4.35–10.91)**	–	16.62 (5.11–54.06)	

Diagnostic ORs for that RfE at the start of a new EoC are given, with CIs in brackets. Significant associations are highlighted in bold type. RfE rubric, ICPC code; RfE label, text label of ICPC code; Nl, the Netherlands; Mt, Malta; Sb, Serbia; Jp, Japan.

TABLE 2 *RfEs with a significant association with the episode title 'depression' (P76), at the start of a new EoC (first encounter in a new episode)—age groups*

		0–14	15–44	45–64	65+
RfE rubric	RfE label	Nl			
P03	Feeling depressed	292.94 (34.70–2472.96)	**121.58 (100.49–147.08)**	**123.10 (97.70–155.09)**	**185.53 (143.33–240.15)**
P01	Feeling anxious/nervous/tense	–	**12.34 (9.38–16.25)**	**13.51 (10.05–18.17)**	**13.01 (8.88–19.07)**
A04	General weakness/tiredness	6.50 (0.80–52.87)	**4.60 (3.59–5.88)**	**3.95 (2.80–5.56)**	**6.12 (4.56–8.21)**
P06	Disturbances of sleep/insomnia	184.04 (43.61–776.61)	**9.05 (6.33–12.94)**	**9.88 (6.81–14.34)**	7.96 (4.78–13.25)
N01	Headache (excl N02 N89 R09)	–	0.53 (0.27–1.08)	0.61 (0.25–1.47)	0.45 (0.11–1.82)
P76	Depressive disorder	9847.00 (1588.01–61 059.51)	**1287.96 (846.30–1960.12)**	**884.70 (570.98–1370.79)**	792.23 (481.26–1304.15)
P04	Feeling/behaving irritable	–	24.26 (13.84–42.54)	24.54 (11.62–51.83)	18.29 (5.58–59.96)
P02	Acute stress/trans/situat distur	–	3.49 (1.55–7.86)	8.15 (4.15–16.00)	10.29 (3.75–28.24)
RfE rubric	RfE label	Mt			
P03	Feeling depressed	–	**402.67 (268.84–603.13)**	160.85 (98.12–263.67)	144.26 (65.64–317.07)
P01	Feeling anxious/nervous/tense	–	**38.73 (26.30–57.03)**	25.43 (15.19–42.56)	27.74 (10.80–71.26)
A04	General weakness/tiredness	–	**6.08 (3.78–9.79)**	2.44 (0.98–6.06)	5.34 (1.84–15.46)
P06	Disturbances of sleep/insomnia	–	58.46 (27.72–123.30)	29.68 (12.73–69.21)	16.99 (3.78–76.49)
N01	Headache (excl N02 N89 R09)	–	1.50 (0.73–3.05)	2.98 (1.37–6.48)	7.71 (2.29–25.98)
P76	Depressive disorder	–	–	–	–
P04	Feeling/behaving irritable	–	57.87 (11.15–300.43)	–	–
P02	Acute stress/trans/situat distur	–	–	–	–
RfE rubric	RfE label	Sb			
P03	Feeling depressed	–	168.70 (46.07–617.69)	110.85 (27.17–452.35)	53.90 (6.18–469.89)
P01	Feeling anxious/nervous/tense	–	10.67 (4.48–25.42)	7.30 (2.90–18.36)	14.28 (5.97–34.11)
A04	General weakness/tiredness	–	2.10 (0.29–15.35)	–	–
P06	Disturbances of sleep/insomnia	–	–	–	–
N01	Headache (excl N02 N89 R09)	–	–	–	–
P76	Depressive disorder	–	–	–	–
P04	Feeling/behaving irritable	–	–	–	–
P02	Acute stress/trans/situat distur	–	34.62 (7.66–156.48)	–	29.94 (3.72–240.82)
RfE rubric	RfE label	Jp			
P03	Feeling depressed	–	–	1316.86 (106.76–16 242.38)	–
P01	Feeling anxious/nervous/tense	–	–	–	–
A04	General weakness/tiredness	–	–	23.26 (5.73–94.38)	5.55 (1.25–24.54)
P06	Disturbances of sleep/insomnia	–	–	12.40 (1.52–101.19)	13.46 (3.02–59.90)
N01	Headache (excl N02 N89 R09)	–	–	3.09 (0.38–24.87)	5.45 (1.23–24.08)
P76	Depressive disorder	–	–	–	–
P04	Feeling/behaving irritable	–	–	–	–
P02	Acute stress/trans/situat distur	–	–	–	–

Diagnostic ORs for that RfE at the start of a new EoC are given, with CIs in brackets. Significant associations are highlighted in bold type. RfE rubric, ICPC code; RfE label, text label of ICPC code; Nl, the Netherlands; Mt, Malta; Sb, Serbia; Jp, Japan. 0–14, 15–44, 45–64, 65+ represent standard age groups.

The ORs and likelihood ratios for the Dutch and Maltese populations are significant, while the ones for Serbia and Japan do not fit our criteria for clinical and statistical significance. This is due to relatively wider CIs, consequent to smaller numbers of observations.

Graphs (years of observation and practices)

The ORs for a diagnosis of depression (P76) given the RfE feeling depressed (P03) at the start of a new EoC in the Netherlands and Malta are illustrated in Figure 1 (none of the ORs from Japan or Serbia fit the significance criteria). The ORs over time (upper graph) and between FDs (lower graph) show good congruency, in that they are all in the same direction away from unity and in many cases are statistically consistent. The ORs in Malta are higher than in the Netherlands even though many do not fit the significance criteria, but in both populations, they appear similar over time. The ORs are consistent between all Maltese FDs (lower graph), despite some practices

TABLE 3 *RfEs with a significant association with the episode title 'depression' (P76), at the start of a new EoC (first encounter in a new episode)—years of observation*

		1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005
RfE rubric	RfE label	Nl										
P03	Feeling depressed	**188.21 (116.67–303.61)**	165.37 (97.92–279.30)	**139.58 (92.92–209.69)**	**157.10 (107.56–229.47)**	**148.33 (96.32–228.41)**	**224.29 (151.62–331.79)**	**134.20 (88.54–203.40)**	**134.26 (91.86–196.23)**	**141.40 (93.94–212.85)**	**198.73 (130.37–302.94)**	87.19 (53.37–142.42)
P01	Feeling anxious/ nervous/tense	24.58 (13.60–44.41)	31.46 (16.90–58.55)	19.69 (12.12–31.98)	13.41 (7.90–22.75)	18.22 (10.68–31.11)	14.00 (8.11–24.18)	10.88 (5.64–21.02)	9.82 (5.11–18.86)	8.14 (3.76–17.60)	14.09 (7.69–25.81)	10.73 (5.18–22.22)
A04	General weakness/ tiredness	9.73 (5.99–15.82)	3.24 (1.48–7.05)	**5.69 (3.65–8.87)**	**6.43 (4.19–9.87)**	5.41 (3.30–8.89)	3.93 (2.30–6.71)	2.49 (1.16–5.35)	3.29 (1.82–5.95)	3.31 (1.68–6.52)	5.77 (3.42–9.75)	3.37 (1.64–6.92)
P06	Disturbances of sleep/ insomnia	14.02 (5.97–32.91)	14.82 (5.84–37.57)	18.47 (10.46–32.62)	19.02 (10.96–32.99)	13.24 (7.06–24.84)	8.23 (3.81–17.81)	11.47 (5.74–22.92)	5.05 (2.05–12.41)	5.29 (1.94–14.43)	5.86 (2.38–14.45)	2.53 (0.62–10.31)
N01	Headache (excl N02 N89 R09)	1.20 (0.29–4.87)	0.72 (0.10–5.20)	0.35 (0.05–2.52)	–	0.34 (0.05–2.41)	0.96 (0.31–3.03)	–	0.36 (0.05–2.61)	1.32 (0.42–4.15)	0.81 (0.20–3.28)	0.47 (0.07–3.37)
P76	Depressive disorder	–	1710.58 (207.93–14 072.70)	1531.12 (676.34–3466.19)	1003.94 (507.52–1985.92)	482.84 (263.98–883.13)	668.85 (318.66–1403.88)	1913.33 (840.28–4356.69)	2273.49 (795.54–6497.16)	755.18 (311.92–1828.35)	1890.50 (651.61–5484.86)	2271.87 (1042.91–4949.05)
P04	Feeling/ behaving irritable	47.21 (15.41–144.65)	53.54 (14.97–191.50)	6.74 (0.91–49.79)	32.02 (9.35–109.70)	21.87 (5.07–94.37)	23.59 (5.41–102.93)	18.98 (2.43–148.06)	–	32.16 (9.55–108.32)	39.47 (15.10–103.17)	15.81 (2.08–120.05)
P02	Acute stress/ trans/situat distur	8.60 (1.14–64.67)	12.71 (3.01–53.69)	6.80 (2.13–21.73)	4.83 (1.18–19.80)	–	10.99 (3.40–35.50)	10.07 (3.12–32.47)	6.72 (1.63–27.67)	–	6.98 (0.95–51.37)	9.60 (2.32–39.82)
RfE rubric	RfE label	Mt										
P03	Feeling depressed							659.09 (336.26–1291.83)	310.74 (165.32–584.09)	442.58 (218.63–895.90)	267.33 (145.93–489.74)	248.11 (131.87–466.78)
P01	Feeling anxious/ nervous/tense							52.06 (26.46–102.40)	52.14 (28.23–96.29)	57.91 (28.08–119.46)	38.16 (19.95–72.98)	31.98 (16.85–60.67)
A04	General weakness/ tiredness							4.90 (1.93–12.43)	7.84 (3.82–16.11)	2.49 (0.60–10.37)	6.82 (3.07–15.13)	7.07 (3.01–16.57)
P06	Disturbances of sleep/ insomnia							83.51 (36.35–191.86)	31.54 (9.08–109.57)	123.09 (41.40–365.96)	–	24.82 (5.53–111.35)
N01	Headache (excl N02 N89 R09)							–	2.15 (0.77–5.96)	3.22 (0.99–10.48)	1.55 (0.38–6.39)	7.26 (3.56–14.80)
P76	Depressive disorder							–	–	268.19 (16.51–4357.56)	–	–

TABLE 3 *Continued*

		1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005
P04	Feeling/ behaving irritable							–	96.64 (8.64–1080.45)	–	–	85.68 (7.68–956.20)
P02	Acute stress/ trans/situat distur							–	–	–	–	–
RfE rubric	RfE label	Sb										
P03	Feeling depressed									118.32 (50.36–278.03)		
P01	Feeling anxious/ nervous/tense									10.39 (6.24–17.30)		
A04	General weakness/ tiredness									0.83 (0.12–5.96)		
P06	Disturbances of sleep/ insomnia									–		
N01	Headache (excl N02 N89 R09)									–		
P76	Depressive disorder									–		
P04	Feeling/ behaving irritable									–		
P02	Acute stress/ trans/situat distur									16.62 (5.11–54.06)		
RfE rubric	RfE label	Jp										
P03	Feeling depressed		–	3766.88 (353.19–40 175.26)	–							
P01	Feeling anxious/ nervous/ tense		–	–	–							
A04	General weakness/ tiredness		50.20 (10.03–251.11)	4.80 (0.61–37.68)	4.62 (0.58–36.74)							

TABLE 3 *Continued*

		1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005
P06	Disturbances of sleep/ insomnia		23.88 (2.74–207.85)	36.38 (7.70–171.85)	–							
N01	Headache (excl N02 N89 R09)		–	5.72 (1.23–26.54)	2.96 (0.37–23.48)							
P76	Depressive disorder		–	–	–							
P04	Feeling/ behaving irritable		–	–	–							
P02	Acute stress/ trans/situat distur		–	–	–							

Diagnostic ORs for that RfE at the start of a new EoC are given, with CIs in brackets. Significant associations are highlighted in bold type. RfE rubric, ICPC code; RfE label, text label of ICPC code; Nl, the Netherlands; Mt, Malta; Sb, Serbia; Jp, Japan.

having wide CIs. In the Dutch population, there seem to be two groups of doctors, three FDs with ORs <100 and three with higher ORs ranging from just <200 to 300.

Four web tables

EoC anxiety disorder and its associated RfEs. Supplementary Table W1 (see online supplementary material) is a similar set of tables for the episode title anxiety disorder (P74). The RfEs with at least one significant association in one population were anxiety disorder (P74, the patient presenting with the disease label as a symptom, as in 'doctor, I have an anxiety disorder') and feeling anxious (P01). There was good international congruence for the RfE P01, although the OR for Japan was not statistically significant. The RfE depression (P74) presented more commonly in the Netherlands. There was good congruence across these tables, especially for the RfE P01 across different age groups, and the ORs in Serbia were significant for all age groups except children. There was also good congruence across years of observation, but ORs were not significant (except in Serbia) in one single year of observation. Very good congruence was observed between FD practices in the Netherlands and Malta, with few exceptions (one FD in Malta had very little data on P74 and one FD in each population had narrower CIs).

EoC tonsillitis and its associated RfEs. Supplementary Table W2 (see online supplementary material) is a similar set of tables for the episode title tonsillitis (R76). Six RfEs were significant in at least one population [in the case of the RfE cough (R05), only in the case of children in Malta]. The clinical concept tonsillitis seems to relate to the same RfEs in all four populations, with good congruence in ORs at population level. 'Throat complaints' (R21), 'fever' (A03) and 'enlarged lymph glands' (B02) seem to increase the likelihood of the diagnosis, while cough (R05) and snuffles (R07) seem to decrease it, and 'pain in the respiratory system' (R01) seems to increase the likelihood only in Serbia. The ORs across age groups and years of observation are congruent, with few exceptions (i.e. mainly in those >65 years, due to small numbers and consequently wide CIs). The conceptualization of the diagnosis seems to occur similarly between FDs in the Netherlands and Malta, again with the phenomenon of wider CIs seen with data from some individual practices and narrower CIs in others.

EoC asthma and its associated RfEs. Supplementary Table W3 (see online supplementary material) is a similar set of tables for the episode title asthma (R96). Five RfEs were significant in at least one of the four populations, with good congruency between the Dutch and Maltese data, similar Japanese data (but with wider CIs) and Serb data exhibiting a wider CI for the RfE

TABLE 4 *RfEs with a significant association with the episode title 'depression' (P76), at the start of a new EoC (first encounter in a new episode)—practices*

		FD1	FD2	FD3	FD4	FD5	FD6
RfE rubric	RfE label	Nl					
P03	Feeling depressed	**304.55 (223.44–415.09)**	**225.60 (182.06–279.54)**	**187.57 (124.38–282.86)**	45.53 (20.03–103.48)	**79.47 (55.21–114.38)**	**55.57 (37.72–81.88)**
P01	Feeling anxious/nervous/tense	**17.89 (12.04–26.58)**	**24.49 (18.58–32.29)**	23.92 (14.06–40.71)	2.92 (0.40–21.15)	3.37 (1.72–6.59)	10.75 (6.18–18.71)
A04	General weakness/tiredness	**9.28 (6.98–12.34)**	**4.86 (3.62–6.53)**	4.98 (3.04–8.17)	6.20 (3.08–12.48)	1.03 (0.46–2.31)	1.72 (0.81–3.66)
P06	Disturbances of sleep/insomnia	**18.60 (12.66–27.32)**	**11.59 (7.89–17.03)**	21.00 (11.00–40.08)	4.59 (1.12–18.88)	2.35 (0.87–6.34)	4.57 (1.68–12.43)
N01	Headache (excl N02 N89 R09)	0.19 (0.03–1.33)	0.72 (0.32–1.61)	1.75 (0.64–4.73)	–	–	0.24 (0.03–1.74)
P76	Depressive disorder	3132.28 (1351.13–7261.46)	637.14 (199.05–2039.41)	1125.23 (137.67–9196.88)	5317.24 (1573.28–17 970.69)	1722.40 (1183.26–2507.19)	714.08 (377.31–1351.46)
P04	Feeling/behaving irritable	25.35 (7.57–84.84)	35.87 (18.38–70.02)	12.47 (2.93–53.07)	–	18.66 (5.67–61.43)	21.55 (8.46–54.90)
P02	Acute stress/trans/situat distur	–	5.63 (2.78–11.43)	12.00 (4.78–30.17)	46.16 (10.32–206.50)	11.30 (2.70–47.28)	4.24 (0.58–30.89)
RfE rubric	RfE label	Mt					
P03	Feeling depressed	366.63 (117.96–1139.48)	305.13 (79.96–1164.44)	528.90 (192.19–1455.52)	**334.48 (241.53–463.20)**		
P01	Feeling anxious/nervous/tense	58.37 (19.11–178.29)	31.31 (7.91–123.92)	–	**50.33 (36.43–69.53)**		
A04	General weakness/tiredness	3.15 (0.41–24.18)	33.73 (9.36–121.61)	3.43 (0.45–25.82)	**5.50 (3.56–8.51)**		
P06	Disturbances of sleep/insomnia	120.42 (29.43–492.82)	21.15 (2.58–173.50)	73.53 (8.74–618.54)	48.02 (25.96–88.82)		
N01	Headache (excl N02 N89 R09)	–	–	–	3.19 (1.96–5.18)		
P76	Depressive disorder	–	–	588.61 (35.43–9778.32)	–		
P04	Feeling/behaving irritable	–	–	–	57.50 (11.10–297.81)		
P02	Acute stress/trans/situat distur	–	–	–	–		

Diagnostic ORs for that RfE at the start of a new EoC are given, with CIs in brackets. Significant associations are highlighted in bold type. RfE rubric, ICPC code; RfE label, text label of ICPC code; Nl, the Netherlands; Mt, Malta; Sb, Serbia; Jp, Japan. FD1–6 represent ORs for an individual FD or FD group practice.

cough (R05). In the Dutch practices, asthmatic patients tended to present with the disease-label RfE bronchitis (R78), a phenomenon not seen in the other populations. The congruency between age groups was good, especially between Dutch and Maltese data, to a lesser degree in the Japanese data, with the Serb data somewhat of an outlier. Excellent congruency was observed across years of observation and practices for most RfEs.

RfE wheezing and associated EoCs. Supplementary Table W4 (see online supplementary material) is a similar set of tables for the RfE wheezing (R03). Four episode titles exhibited significant associations with this RfE: bronchitis (R78), head cold (R74), asthma (R96) [to be expected from the data in Supplementary Table W3 (see online supplementary material) above] and tracheitis (R77). There is very good congruency in diagnostic ORs among the four populations, with the exception of the episode title head cold (R74) having a significant OR in the Dutch population but not in the other three. Congruency is also evident across age groups, years of observation and practices, while the previously observed effect of widening CIs with smaller numbers did not allow us to describe even more associations. Serb FDs do not seem to manage children presenting with the RfE wheezing (R03) or asthma and tracheitis in those >45 years of age.

Discussion

Summary

First research question. This paper quantifies exemplar relationships between common RfEs and common diagnoses (episode titles) in practice populations from

TABLE 5 *ORs, likelihood ratios, sensitivity, specificity, positive and negative predictive value pre- and post-test odds for the episode title 'depression' (P76) given the RfE 'feeling depressed' (P03) at the beginning of a new EoCs in the Netherlands (Nl), Malta (Mt), Serbia (Sb) and Japan (Jp)*

P03 (feeling depressed)	Depression (P76)				
	Nl	Mt	Sb	Jp	Formula
(*N*) With P76 and RfE P03	461	135	8	5	a
(*N*) With P03 and other EoC	895	136	17	1	b
(*N*) With P76 and other RfE	1132	158	163	22	c
(*N*) Without P03 and other EoC	334 860	55 392	40 984	22 770	d
OR (CI)	**152.37 (134.25–172.93)**	**348.00 (261.77–462.65)**	118.32 (50.36–278.03)	5175.00 (580.68–46 119.54)	ad/bc
LR+ (CI)	**108.56 (98.13–120.10)**	**188.12 (152.70–231.76)**	112.83 (49.36–257.94)	4216.85 (509.39–34 908.41)	Sens/1-Spec
LR– (CI)	0.71 (0.69–0.74)	0.54 (0.49–0.60)	0.95 (0.92–0.99)	0.81 (0.68–0.98)	1-Sens/Spec
Sens	0.29	0.46	0.05	0.19	a/a + c
Spec	1.00	1.00	1.00	1.00	d/b + d
PV+	0.34	0.50	0.32	0.83	a/a + b
PV–	1.00	1.00	1.00	1.00	d/c + d
Pre-test odds	0.00	0.01	0.00	0.00	Prev/1-Prev
Post-test odds	0.52	0.99	0.47	5.00	Pre-test odds × LR+
Prevalence	0.00	0.01	0.00	0.00	a + c/a + b + c + d

The numbers of cases of depression and other episode titles with and without the RfE 'P03' are given. Diagnostic ORs for that RfE at the start of a new EoC are given, with CIs in brackets. Significant associations are highlighted in bold type. Prevalence reflects the prevalence rate in a population of EoCs, and not in a population of patients. LR+, positive likelihood ratio; an expression of the extent to which a symptom increases the probability of a diagnosis. The likelihood ratio for the existence of the symptom (RfE) is the odds that it will exist in a new EoC with that diagnosis, in contrast to a new EoC without that diagnosis; LR–, negative likelihood ratio; the likelihood ratio for absence of the symptom (a negative result) is the odds that a test will be negative in a new EoC of that diagnosis, contrasted with an EoC without that diagnosis; PV: predictive value (+, positive and –, negative); the probability that a new EoC with a positive test (presence of a defined RfE) has the disease (positive predictive value). The probability that a person or a proportion of a population with a negative test does not have the disease is the negative predictive value; odds (OR): diagnostic OR of disease (i.e. odds of episode title P74 present against absent) against test (i.e. odds of RfE P03 present against absent); the ratio of the probability of occurrence of an event to that of non-occurrence; Sens, sensitivity; a test with high sensitivity detects a high proportion of true cases; Spec, specificity; the specificity is the proportion of truly non-diseased persons who are so identified by the test (synonym: true negative rate); Pre-test: pre-test odds; odds of disease in all new EoCs; Post-test: post-test odds; odds of disease in the population of new EoCs starting with the RfE P03. Note that programme curtails post-test odds to 0.99 if ⩾0.99.

Malta, the Netherlands, Japan and Serbia, across age groups, periods of observation and practices.

Second research question. The generic similarities in the relationships between common RfEs and common diagnoses (episode titles) in these practice populations appear to be far more remarkable than any differences. In fact, hardly any significant examples of the latter were found. There was marked congruence in the direction of diagnostic relationships between populations, across age groups, years of observation and practices. As expected, we found variability in the magnitude of such associations, rather more across age groups and between countries rather less between years of observation and practices.

Despite differences in magnitude of ORs between populations, virtually all which fit our significance criteria (besides many which did not) were in the same direction from unity. Some diagnostic associations, especially in the Serb and Japanese populations, may have been too infrequent to estimate precisely and consequently did not fit our intentionally conservative significance limits.

These findings support the aggregation of FM diagnostic data across age groups, years of observation and practices.

Third research question. There seems to be a strong common trend in the diagnostic associations we found, especially in their common direction away from unity in different observation frames. In our view, these similarities reflect common core diagnostic concepts and processes and support the existence of an international core diagnostic process in the domain of FM. On this basis, we would support the utility of diagnostic ORs from one population applied to another but would recommend country-specific data where available.

Analysis of exemplar diagnostic associations

Depression. For the episode title depression, the similarity between the Dutch and Maltese diagnostic ORs at population level is striking, with few differences observed (Table 1). Even then, such differences are minor: for example, in the case of the RfE sleep disturbance (P06), the Maltese OR is clinically but just not statistically significant, and the large Japanese OR has a wide CI, but all three ORs are in the same direction from unity. Practically, all the other ORs were in the same direction from unity. The Serb and Japanese ORs appear to differ from the other two sets only in the sense that no one

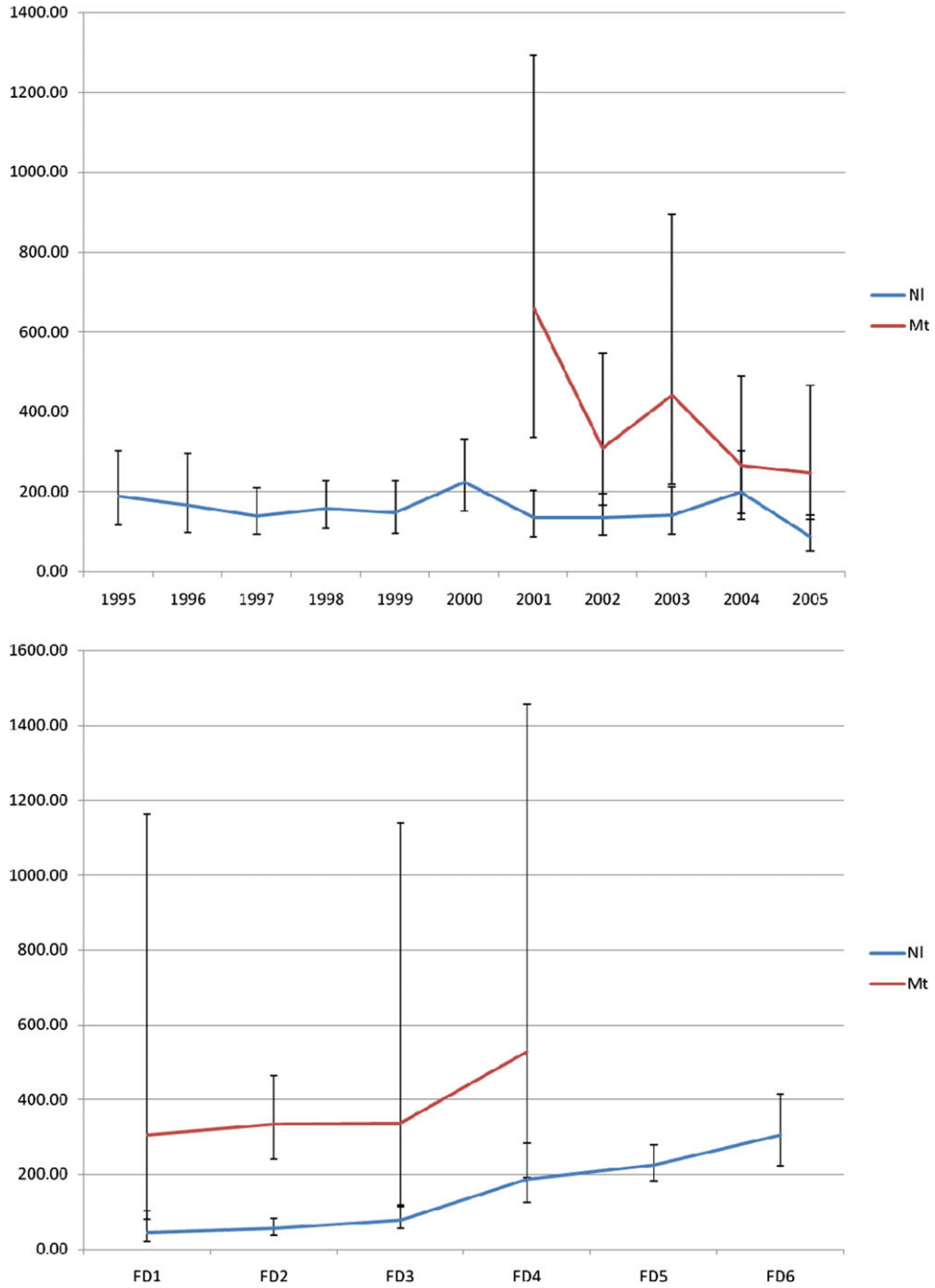

FIGURE 1 *Diagnostic ORs for depression (P74) given the RfE sadness (P03) at the start of new EoCs in the Netherlands and Malta. ORs by year of observation (top graph) and by practice (bottom graph). X-axis gives year of observation (top graph) and practice number (from FD1 to FD6, four practices in Malta and six in the Netherlands); Y-axis gives the OR. The two traces represent data from the two populations from Malta and the Netherlands. The 'whiskers' give the 95% confidence limits for the OR (CI)*

diagnostic association is both clinically and statistically significant. The observation window is evidently too narrow to obtain reliable data given the prevalence of the condition in these two populations (1 year in Serbia and 3 years in Japan). Rather than confirm any difference, we could not confirm similarities. Larger databases would have allowed more reliable estimates and more significant associations to be studied.

We found some minor differences across age groups (Table 2). ORs seem to be congruent (similar in size, but not necessarily overlapping CIs) across the age groups from 15 years of age upward for most RfEs in the Netherlands, with wider CIs in children. However, for the RfEs 'feeling irritable' (P04) and 'stress' (P02), the CIs are simply too wide to confirm similarity, while headache (N01) does not contribute. Similarly, in Malta, the first four RfEs are congruent across all age

groups, except for children and except for headache (N01) in the 15–44 age group (CI includes unity), but statistically significant ORs were only found in the 15–44 age group. Statistically significant ORs for depression were not observed in Serbia and Japan, but still those for three RfEs seemed similar across age groups and to those from the other two populations, especially in direction away from unity.

We found high degrees of congruency between diagnostic ORs across years of observation (Table 3). Again, differences were due to lack of power and small numbers of observations rather than the confirmation of any clinically significant differences. For example, the RfEs feeling depressed (P03) and tiredness (A04) appear to be clinically but just not statistically significant (the CIs are just too wide) when comparing single years of observation from Malta.

We only found small differences between FD practices (Table 4). The Dutch data exhibit good congruency for all eight RfEs, with similar ORs for seven RfEs and no significant association with the RfE headache (N01). Headache (N01) did not contribute to diagnosing depression (P76) in our Dutch population, either at practice or at population level (compare Table 1). All the ORs from the different Dutch practices were in the same direction from unity, with different reliability (CI width); for example, for FDs Numbers 5 and 6 and for the RfE 'tiredness' (A04), many CIs included unity. In Malta, we also found good congruency across practices, but one practice had relatively narrower CIs due to increased workload (data not tabulated). We found similar clinically significant ORs in four of eight RfEs, but most Maltese practice data were not powerful enough to achieve statistical significance. For the other four RfEs, we did not have enough data to comment on similarities or differences, either at practice or at population level (compare Table 1). Reassuringly, a clinically significant association with the RfE headache (N01) which was almost statistically significant in one practice was also just statistically significant at population level (see Table 1). In general, ORs between practices exhibit good congruence, but some observations have wider CIs, and consequently, clinically significant associations may not achieve statistical significance. Data from the individual practices reflected the aggregated population data rather well but lacked the power to reliably estimate some clinically significant associations.

Summarizing, we found the content of the diagnostic concept depression, as defined by relationships with RfEs presenting in FM, to be very similar in these four populations, across age groups, years of observation and practices. There were some small differences across age groups, especially children, and between populations, in magnitude rather than direction of these diagnostic associations. We could not reliably define diagnostic relationships (ORs) in some cells due to insufficient power of the data, especially when split in this way. This is also a reflection of lower prevalence of the condition, and consequently less experience and expertise, in Serbia [1] and Japan, and also in some individual FD practices. The wide CIs of many ORs do not allow us to confidently conclude that the diagnostic concept is identical in four countries, but certainly do not exclude such. Although it is possible that the diagnosis of depression is conceptualized differently in these four populations, the data suggest otherwise. In fact, we found more similarities than differences. Aggregating data across age groups, years of observation and practices adds statistical power, and the limited variability we have described supports such aggregation.

Anxiety

Similar conclusions can be drawn for the other diagnostic associations studied (data in web-based tables). In the case of anxiety [P74, supplementary Table W1 (see online supplementary material)], the ORs for feeling anxious (P01) as an RfE were congruent for three countries, and clinically but not statistically significant in Japan. Serb FDs experienced higher exposure (higher prevalence)[1] and consequently, 1 year of observation was sufficient for reliable data for clinically significant ORs. Splitting of the data in various cells resulted in wide CIs for the clinically significant relationship with 'P01' in many cases, but the ORs were in the same direction in most contrasts. It is notable that although some FDs had narrower CIs than others for an OR, ORs were very similar for the RfE feeling anxious (P01) for both Dutch and Maltese FDs.

Tonsillitis

The diagnostic concept of tonsillitis [R76, supplementary Table W2 (see online supplementary material)] also seems to be very similar among the four populations, with ORs for the significant RfEs being in the same direction from unity, with the exception of the RfE respiratory system pain (R01) in Serbia. One suspects that this latter observation may be due to coding of throat pain (R21) with a less specific code of respiratory system pain (R01) by Serb FDs. The high prevalence of the condition in all four populations allows reliable estimation of more clinically significant diagnostic ORs. The increased power of the data allows us to conclude that the clinical concept is remarkable similar in all four populations: throat pain, fever and enlarged lymph glands contribute to making the diagnosis of tonsillitis, while cough and sneezing contribute to excluding it. This 'clinical picture' is consistent between populations and individual practices, across age groups and years of observation. The widening or narrowing of CIs in the case of individual practices reflects effects of differences in workload and exposure. Nonetheless, the ORs are in the same direction from unity, practically without exception. The reliably estimated ORs defining the symptomatology of

tonsillitis are both a reflection of the expertise and experience of individual FDs and evidence of a common underlying clinical concept crossing national divides.[2]

Asthma and wheezing

Good congruency is also seen for the episode label asthma (R96) and the RfE wheezing (R03) [supplementary Tables W3 and W4 (see online supplementary material), respectively], reflecting common clinical concepts. The wide CI of many ORs from Serbia and Japan reflects that a narrow observation period (3 years in Japan and 1 year in Serbia) is not enough to obtain reliable ORs in different age groups, especially for clinical conditions of lower prevalence. Apparently, children are seen by community paediatricians in Serbia, and this is reflected in the wider CIs or empty cells for this age group. The data from Malta and the Netherlands allow more in-depth interpretation. However, even in these two populations, it is evident that less clinical exposure and lower observed frequency are reflected in the data with wider CIs for the observed ORs, as we have described previously.[2]

Age groups

Aggregation of data across age groups adds power, although the availability of OR data for different age groups is useful. One notes that the expertise of diagnosing depression in children is limited in all four populations, with no significant associations in this age group. Serbia in particular provided very little data on children across the board.

Period of observation

The congruence in ORs over time, also illustrated in Figure 1, supports the pooling of data from different years of observation. The CIs widen with less available data, and 1 year of observation is not powerful enough for less common conditions. The larger OR with a wider CI from Malta in 2001 is consequent to a smaller population in the first year of observation since this population is based on patients consulting. Again, the aggregation of data across years of observation adds power and is recommended by this study, considering the relative lack of observed variability over time.

Practices

Inter-doctor variation of diagnostic associations was not marked and diagnostic ORs were congruent between individual FD practices in our study. Limited variability was noted, in the sense that some diagnostic ORs were larger than others or might have wider CIs, sometimes including unity. We did not find significant diagnostic associations in a different direction from unity between FD practices, with remarkably few exceptions.

If 'only' one FD practice picked up (or failed to pick up) an association in isolation, the effect on the pooled country OR was not sufficient to reach significance (or conversely to not become significant) on the basis of the data from one single practice. Examples include the weak association between the RfEs headache (N01) and irritability (P04) and the EoC depression (P76) in Malta and individual Dutch practices not finding a significant association between anxiety (P01) or tiredness (A04) and depression (P76).

These findings are in agreement with previous studies which show that the cluster effect in family medicine is small, with the exception of differences in process distributions.[3,4] One expects to observe variance in diagnostic associations between doctors. However, although some FDs have more experience than others, this variance is small and the congruence of findings is remarkable. Again, this supports aggregating data across practices.

Similarities and differences

The observation that almost all the significant diagnostic ORs were in same direction around unity for all RfE to episode title relationships in the four populations, across age groups, years of observation and between practices, in all five clinical scenarios, is both highly remarkable and reassuring. Clinical concepts, as defined by these associations, seemed very similar between populations, even though the strength of the association between a symptom and a diagnosis might have varied to a limited degree. Country-specific data are desirable, but diagnostic data from one population could be used on another, with limitations.

The results of this study support the aggregation of such diagnostic data across observation frames. It also supports the existence of an international body of medical knowledge which is common to FDs in different populations, even though it is not conclusive proof of such. This latter hypothesis has not been tested previously, to our knowledge.

Diagnostic expertise

We propose that lower prevalence of a specific disease, lower levels of exposure to its related diagnostic challenges, less FD experience and less FD expertise go hand in hand. Consequently, one cannot obtain reliable data on a diagnostic association without sufficient exposure of FDs to the same and sufficient data from their practice.[2] On the other hand, exposure alone, without expertise, would give reliable data which would demonstrate variable diagnostic performance between FDs.

Diagnostic ORs reflect not only FD exposure to RfEs and episode titles but also to their diagnostic associations. As such, the expertise of FDs is indeed reflected in appropriately picking up and recording these diagnostic associations. The more such associations are appropriately recorded, the more observations are available to produce reliable data on such ORs and the more congruency and consistency is demonstrated in comparisons such as the ones we have published.

We have indeed found reliable evidence of consistent diagnostic performance from FDs with high exposure and, conversely, did not find any evidence of variable diagnostic performance and inconsistent diagnostic associations, neither in this paper nor in an earlier study in the series.[2] We feel that these observations give evidence that exposure, experience and expertise go hand in hand and that such is reflected mathematically in our analyses.

Clinical and statistical significance limits

The data for the diagnostic OR calculations for the RfE feeling depressed (P03) and the episode title depression (P76) in the four populations are listed in Table 5. These calculations exemplify the large ORs (and LRs) analysed in this study, which are due to the large denominator ('d' in the equation OR = a × d/b × c) provided by large populations under study for long observation periods. The narrowest CIs are found in the Dutch and Maltese datasets, also due to comparatively more observations. In the case presented in Table 5, it is only the LR+ for 'sadness' (P03) and depression (P76) which were clinically significant in two of four populations. No population LR– was clinically significant. Thus, the RfE 'P03' makes a positive contribution to the diagnosis of 'P76', but its absence does not help one to exclude that diagnosis. The OR summarizes this information in one number. The literature suggests that the arbitrary cut-off for a clinically significant LR+ should be ⩾2.0 and that for an LR– should be ⩽0.5.[12–20] We have chosen more conservative cut-offs due to our use of ORs rather than relative risk ratios (see Methodology) because the OR summarizes both the LR+ and LR– and its significance limits should not simply be the same as either of the latter two.

Are such clinical significance limits too conservative? Previous studies of the process of diagnosis [12–20] show what a small effect an LR+ of 2 would have on increasing the posterior (post-test) probability of an index condition.[10] The decision is ultimately arbitrary, and we respect choices different from our own. Our criteria for an OR (⩾3, ⩽0.3, with a CI less wide than the size of the observation itself) allow one to be more confident of including such a datum in a clinical prediction rule since it would have a larger effect.

It is easier to defend excluding ORs, which are clinically insignificant according to our criteria, since these do not contribute much to making or excluding a diagnosis. On the other hand, the temptation to include clinically significant ORs which may have a wide CI, but which do not include unity, is strong. Wide CIs for an OR which is clinically significant are due to small numbers of observations and lower power. In these cases, there is evidence of an association, but it is not reliable. Giving in to this 'temptation' would involve loosening the statistical significance criteria and would include a number of diagnostic ORs based on a small number of observations. The generalizability of such data is suspect. It is arguably a more clinically sound approach to recommend the use of data which fit our stricter criteria for clinical and statistical significance and which are consequently based on more observations. This approach also statistically adjusts for making multiple comparisons.[11]

We observed a rapid widening of CIs as we examined cases with less data (less frequent conditions and RfEs, from populations with smaller observation windows). One indeed needs large datasets to obtain reliable ORs for less common diagnostic associations,[21] and in fact, very large datasets to obtain reliable ORs for less common associations in individual age groups. Rather than loosening one's significance criteria, it is advisable to improve precision by aggregating data from a larger number of FDs and over a longer observation period. Such pooling of data is well supported by the observations we have described above. It is evidently better to combine ORs as we have done previously[2] than to loosen significance criteria to accept observations based on small numbers of observations. With our tightened statistical significance criteria, we adjust for the limited effects of clustering of data.[11]

It is important to add at this stage that one may combine LRs for different RfEs, besides other clinical data such as test results, to form a clinical prediction rule based on multiple predictors. In this case, combining a number of small LR+s (say three RfEs with an LR+ of 1.5) may together produce an appreciable effect (1.5, times 1.5, times 1.5, is 3.4), but inevitably, the CI for this combined LR+ will be widened substantially. One must note that combining LRs in this way may break the rule of conditional independence of observations required for Bayesian analysis.

Our considered recommendation is to have strict criteria for clinical and statistical significance, but not to go so far as to expect diagnostic associations to be statistically consistent with overlapping CIs. The narrower the CIs, the more unlikely that any pair of ORs will be statistically consistent. We recommend that if a set of diagnostic associations are reliably found to be in same direction away from unity and clinically and statistically significant, such associations are to be accepted as congruent. Congruent diagnostic associations such as these would be a sound basis for a clinical prediction rule or a diagnostic guideline.

Validity

Clinical decisions, including diagnostic decisions, should ideally be supported by evidence. Such evidence should be based on empirical studies of clinical practice. Unfortunately, empirical data on diagnosis in FM are currently lacking. The data we present in this paper are based on actual clinical practice, but one may question whether such data really represent an evidence base for

best practice. However, if one does not use such empirical data, then what evidence should one use?

This paradox presents itself to the researcher in the field of diagnosis in FM: a diagnostic decision not based on an evidence-based guideline may be questioned, but one could also question the validity of diagnostic guidelines based on evidence from other domains, such as secondary care. This paradox will continue endlessly unless we address the current lack of evidence for diagnostic decisions in FM. We present these data to hopefully start a trend to address this lack of evidence.

The data from one single practice will hardly ever be enough to provide evidence for a diagnostic association, besides the fact that such data would not be generalizable. Another argument therefore exists for combining ORs across practices and years of observation, and this is the general approach in published studies on diagnostic relationships.[9,13–20]

We externally validated our empirical diagnostic model for depression. The *Diagnostic Statistical Manual of Mental Disorders, 4th edition (DSM-IV)* criteria for depression include the symptoms: feeling depressed or sad, anhedonia, appetite or body weight changes, sleep disturbance, agitation or restlessness, lethargy or tiredness, feelings of worthlessness or guilt, problems with concentration and thinking and thoughts of self-harm.[22] We found the following symptom associations in at least one population in our study: feeling depressed (P03), feeling anxious/nervous/tense (P01), general weakness/tiredness (A04), disturbances of sleep/insomnia (P06), headache (N01), depressive disorder (as an RfE, P76), feeling/behaving irritable (P04) and acute stress (P02). We did not find an association for the RfEs 'weight change' and 'appetite change', listed in the DSM-IV criteria, either because we did not have enough data to estimate a reliable OR or because such an association is not clinically significant in the less severe cases seen in primary care. Additionally, such symptoms may possibly not present at the earliest stages of clinical depression, when we have measured associations at the beginning of an EoC. Other symptoms and signs may not be common enough to have their own ICPC code, such as 'feelings of worthlessness'. With these plausible exceptions, the correlation between the DSM-IV criteria for this diagnosis and the RfEs we observed empirically, is indeed remarkable. We consider that this external validation of our data was supportive of our study and our model.

Limitations

Our analyses were limited to only four EoCs and one RfE, and one may challenge our broad conclusions on this basis. However, in our study of the data from the Transition Project over the past years, we have indeed found many similar trends. We are confident that what we have presented is quite typical of the distributions of diagnostic associations in these populations.

The observed variation in diagnostic associations between and within populations is one manifestation of a complex adaptive system, being subject to multiple interacting effects (e.g. geographical location, demography, culture, socio-economic effects, co-morbidity, inter-doctor variation, changing medical practice over time, etc.). It is not possible to tease out such individual effects in a complex adaptive system such as is the practice of FM.

The diagnostic ORs and models presented in this paper are limited in that they represent an analysis of diagnostic associations at the start of an EoC (first encounter for a new episode) and do not take into account that the diagnosis may have changed later during the episode. Such data are captured in the Transition Project and will be the subject of a planned future study. Furthermore, it is quite possible to miss rare, but important, diagnostic associations due to their infrequent nature and the wide CI for such an association. These data guide but do not replace the expertise of an experienced FD.

FDs are often selected to participate in EMR research projects after they have voluntarily accepted to record such data. Thus, such FDs are often not representative of all FDs in a national system but rather tend to collect data at a higher level of detail and accuracy than their colleagues. They may also receive an incentive to do so (financial or academic). Thus, the analysis of such data exhibits many of the qualities and limitations of both qualitative and quantitative research methodologies, sacrificing some generalizability for increasing depth and accepting inherent biases which cannot be adjusted for without introducing new systematic error.

The ORs we present in this paper represent statistics calculated from a number of practices, and they are not corrected for the effect of clustering. However, such effects are small, and our criteria for considering an OR as significant were tightened to avoid type 1 error.[11]

The analyses we performed were one way for single RfE and episode title combinations. The widening of CIs with smaller numbers, and the rapid decay as one drills down to ORs for separate age groups or for less common RfEs, implies that enormous datasets are required for studying anything but the most prevalent diagnostic associations. This also poses a significant challenge for multivariate models, which combine data from a number of RfEs to predict the likelihood of a diagnosis in a population, especially one for a defined age group. Another methodological challenge is that RfEs are not independent observations since they occur together in one encounter and thus violate the condition of independent observation required by many statistical procedures. Various methodological solutions exist which allow one to correct for failure of conditional independence or to correct for the effect of clustering or for the interactions of multiple variables. However, these models would require very large samples to allow such iterative computations, which rapidly become very

complex with the inclusion of even a few variables. Producing a more precise data model to predict a diagnostic outcome given a set of variables is indeed a significant challenge, which shall be probably only met with large datasets involving data pooled from a large group of FDs, over a long period of time.

Strengths

This study reports an original international comparative analysis of the relationships between RfEs and episode titles during routine FM care of practice populations, including a study of variation in diagnostic approach and diagnostic concepts in four populations, across age groups, years of observation and practices. It is one of few such studies in the domain of FM.

The data analysed in this research project allow the study of relationships between diagnoses and RfEs and the calculation of posterior (post-test) probabilities for a diagnosis in an individual with a defined symptom. This allows for the study of such Bayesian probabilities and functions (prior probability, likelihood function, posterior probability) for practically all combinations of RfEs and episode labels coded in ICPC. This is the first such research project to publish such data within an international comparison. These data are of value for decision support systems to support diagnosis in primary care.

The diagnostic entity is often 'forced' into a disease-label diagnosis, even when it does not entirely fit the diagnostic criteria. This is an anomaly often found with the application of such classifications which are not primary care oriented, such as the International Classification of Disease, and which do not facilitate labelling symptom diagnoses and the efficient handling of diagnostic uncertainty. With ICPC, the availability of the symptom diagnosis keeps disease-label diagnostic classes (rubrics) clean.[5] The similarities we have observed are therefore reinforced in their validity.

Implications

This study has explored a number of important aspects of the use of EMR data for FM research. We have continued to develop the methodology for studying diagnostic associations with ICPC in an international comparison of EoCs in daily family practice, following up on previous articles in this series.[1,2] We have expanded on these studies by looking at the effects of age, time window and individual practices and came to the conclusion that EMR data can be usefully aggregated from different observation and sampling frames. We have considered the effects of such aggregation on the reduction of data complexity. We have made some suggestions on clinical and statistical significance limits appropriate for looking at large numbers of diagnostic associations and considered the effects of small numbers of observations as one 'drills down' to individual practice or age group data. All these issues will hopefully inform future research on EMR data from FM.

Finally, our informed reflections on the effects of FD expertise on diagnostic performance and on the international body of medical knowledge shared among FDs in different countries are proposed as intriguing observations to be further tested.

Conclusions

We confirmed our earlier findings and found little variability of diagnostic associations, especially across years of observation and between individual FD practices. We found some variability of diagnostic associations across age groups and between populations in different countries.

There is a lot of congruence in diagnostic concepts in the domain of FM between populations, across age groups, years of observation and practices. The strength and distribution of these diagnostic associations are not equal between populations, but those reliably estimated were in the same direction in virtually all cases. More data would have allowed more power to define more of such diagnostic associations but would unlikely have changed our conclusions. We found evidence to support an international core diagnostic process in FM. The main conclusion of this study is that we can, and should, aggregate data from different practices and across years of observation in a population. We estimate the cluster effect in FM to be a relatively small contributor to observed variability of diagnostic associations.

We recommend clinical and statistical significance limits as follows: an LR+ of $\geqslant 2$, an LR– of $\leqslant 0.5$, a diagnostic OR of $\geqslant 3$ (2.45) or $\leqslant 0.3$ (0.34), each with a CI which is less wide than the size of the observation itself. Such limits help to avoid making clinical recommendations on very small numbers and adjust for the clustering effect in aggregated data.

Declaration

Funding: The European Union Financial Protocol 7 project 'TRANSFoRm' (www.transformproject.eu, FP7 247787) supported part of the protected time of the authors in performing this study, through its partner the Mediterranean Institute of Primary Care (www.mipc.org.mt).

Ethical approval: The study did not involve the collection of new data. Ethical approval was applied for locally, when appropriate, for individual studies based on these data in the Netherlands, Serbia Malta and Japan.

Conflict of interest: none.

Supplementary material

Supplementary material is available at *Family Practice* online.

Acknowledgements

Author contributions: JKS developed the research methodology, collected data (from Malta), analysed data, was the lead writer of the manuscript; IO developed the research methodology, analysed data, participated in writing the manuscript; SO developed the research methodology and analysed data; KvB developed the research methodology and collected data; PZ collected data; MJ collected data; FD developed the research methodology and participated in writing the manuscript; HL developed the research methodology, collected data, analysed data, participated in writing the manuscript until his death in late 2008. The authors would like to thank the associate editor, Prof. Martin Dawes, and the three reviewers who helped improve this paper with their detailed review and feedback. This study would not have been possible without the participation of the Transition Project doctors. From the Netherlands: C. van Boven MD, PhD, Franeker; P.H. Dijksterhuis MD, PhD, Wirdum and Olst; A. Groen, MD, Amstelveen; J. de Haan, MD, Franeker; A.M.Honselaar-De Groot MD, Amstelveen; D. Janssen MD, Franeker; T.A.L. Polman MD, Franeker; G.O. Polderman MD, Amstelveen; K.E.I. Stolp MD, Amstelveen; N. Valken MD, Wirdum; M.T.M. Veltman MD, PhD (deceased), Amstelveen; M. Woerdeman MD, Amstelveen. From Malta: Francis Paul Calleja MD, Birkirkara; Carmen Sammut MD, Siggiewi; Mario R Sammut MD MSc, Siggiewi; Daniel Sammut MD, Zabbar; David Sammut MD, Zabbar; Jason Bonnici MD, Zabbar; John Buhagiar MD, Zabbar; Andrew Baldacchino MD, Zabbar. From Serbia: the FDs in the region of Kraljevo, part of the ICRC project.

References

[1] Soler JK, Okkes I, Oskam S *et al*. An international comparative family medicine study of the Transition Project data from the Netherlands, Malta and Serbia. Is family medicine an international discipline? Comparing incidence and prevalence rates of reasons for encounter and diagnostic titles of episodes of care across populations. *Fam Pract* 2012; **29:** 283–98.

[2] Soler JK, Okkes I, Oskam S *et al*. Is family medicine an international discipline? An international comparative family medicine study of the Transition project data from the Netherlands, Malta and Serbia. Comparing diagnostic odds ratios across populations. *Fam Pract* 2012; **29:** 299–314.

[3] Marinus AMF. Inter-doktervariatie in de huisartspraktijk. *PhD Thesis*. Amsterdam, The Netherlands: University of Amsterdam, 1993.

[4] Boerma WGW. *Profiles of General Practice in Europe: An International Study of Variation in the Tasks of General Practitioners.* Utrecht, The Netherlands: Nivel, 2003.

[5] Soler JK, Okkes I, Lamberts H, Wood M. The coming of age of ICPC: celebrating the 21st birthday of the International Classification of Primary Care. *Fam Pract* 2008; **25:** 312–7.

[6] http://www.transitieproject.nl (accessed on 1 November 2009).

[7] Wonca International Classification Committee. ICPC-2: International Classification of Primary Care, 2nd edn. Prepared by the International Classification Committee of WONCA (WICC). Oxford, UK: Oxford University Press, 1998.

[8] Okkes IM, Oskam SK, Van Boven K, Lamberts H. EFP. Episodes of Care in Family Practice. Epidemiological data based on the routine use of the International Classification of Primary Care (ICPC) in the Transition Project of the Academic Medical Center/University of Amsterdam. (1985–2003). In: Okkes IM, Oskam SK, Lamberts H (eds). *ICPC in the Amsterdam Transition Project*, CD-Rom. Amsterdam, The Netherlands: Academic Medical Center/University of Amsterdam, Department of Family Medicine, 2005.

[9] Okkes IM, Oskam SK, Lamberts H. The probability of specific diagnoses for patients presenting with common symptoms to Dutch family physicians. *J Fam Pract* 2002; **51:** 31–6.

[10] Feinstein AR. Indexes of contrast and quantitative significance for comparisons of two groups. *Stat Med* 1999; **18:** 2557–81.

[11] Altman DG, Machin D, Bryant T, Gardner MJ. *Statistics with Confidence. Confidence Intervals and Statistical Guidelines.* BMJ Books, 2000.

[12] Jaeschke R, Guyatt GH, Sackett DL. User's guides to the medical literature III. How to use an article about a diagnostic test. Are the results of the study valid? *JAMA* 1994; **271:** 703–7.

[13] Metlay JP, Kapoor WN, Fine MJ. Does this patient have community-acquired pneumonia? Diagnosing pneumonia by history and physical examination. *JAMA* 1997; **278:** 1440–5.

[14] Bundy DG, Byerley JS, Liles A *et al*. Does this child have appendicitis? *JAMA* 2007; **298:** 438–51.

[15] Cass SA, Vollenweider MA, Hornung CA, Simel DL, McKinney WP. Does this patient have influenza? *JAMA* 2005; **293:** 987–97.

[16] Bent S, Nallamothu BK, Simel DL, Fihn SD, Saint S. Does this woman have an acute uncomplicated urinary tract infection? *JAMA* 2002; **287:** 2701–10.

[17] Shaikh N, Morone NE, Lopez J *et al*. Does this child have a urinary tract infection? *JAMA* 2007; **298:** 2895–904.

[18] Margaretten ME, Kohlwes J, Moore D, Bent S. Does this adult patient have septic arthritis? *JAMA* 2007; **297:** 1478–88.

[19] Trowbridge RL, Rutkowski NK, Shojania KG. Does this patient have acute cholecystitis? *JAMA* 2003; **289:** 80–6.

[20] Ebell MH, Smith MA, Barry HC, Ives K, Carey M. The rational clinical examination. Does this patient have strep throat? *JAMA* 2000; **284:** 2912–8.

[21] Buntinx F, Aertgeerts B, Aerts M *et al*. Multivariable analysis in diagnostic accuracy studies: what are the possibilities?. In: Knottnerus JA, Buntinx F (eds). 2nd edn. *The Evidence Base of Clinical Diagnosis. Theory and Methods of Diagnostic Research*, Oxford, UK: John Wiley & Sons Ltd, 2009.

[22] *Diagnostic and Statistical Manual of Mental Disorders (4th edn)*. Washington, DC, American Psychiatric Association, 1995.