

The University of Nottingham

Extracting epidemiological data from electronic medical record recording systems in United Kingdom Primary Care

Vishal Rai Basra, BSc MSc

Thesis submitted to the University of Nottingham for the degree of Master of Philosophy

August 2017

Abstract

Electronic medical records (EMRs) in primary care represent a longitudinal record of patient's current and past history and important health related data. The ability to use of electronic medical records from a variety of different sources is an important and growing area of research. This study aims to use participants in an existing cohort of patients (the Gedling study) and extract their GP health data into a usable research database.

The General Practitioners of each study participant were contacted and provided with evidence of consent to extract the records. The software packages in use at participating practices were examined to determine the most suitable format for data export, and the process documented. Various methods were considered to convert the extracted records into a research-suitable format, with the use of regular expressions being selected. Comparisons were made between the amount and type of data recorded among the software packages. The content of the converted medical records was compared with the information obtained directly from participants through the Gedling study surveys.

There were notable differences across different GP software packages; particularly with regard to the amount of free text which is recorded in individual patient records with the most free text found in older patients, women and more deprived groups. There were some notable differences in the numbers of prescriptions and consultations recorded in each package. Self-reported asthma, wheeze, COPD cases, and hay fever recordings showed fair agreement with medical records.

EMRs are increasingly the main mechanism used to record patient data and are fast becoming a large repository of clinical information. Exploring this potential it can contribute to our understanding of patient needs, prescribing patterns,

accurate prevalence and incidence of disease and whether national policies for healthcare are delivering the benefits they set out to achieve.

Acknowledgements

The work in this thesis was conducted under the supervision of Dr Tricia McKeever and Dr Jack Gibson at the University of Nottingham. It was funded by the Economic and Social Research Council (ESRC) studentship programme.

I would like to express my deep gratitude and thanks to my supervisors for their support and guidance.

My thanks to all the PhD students in B126 and to all my other friends along the journey.

Thanks to my Mum and Dad and all my family for their encouragement and support.

I would like to dedicate this thesis to my close friend who encouraged me to apply in the first but passed away just prior to his own PhD graduation ceremony. This is dedicated to the memory of Dr Nick Williams.

Table of Contents

Abstract i
Acknowledgements iii
List of Tables x
List of Figures xi
List of abbreviations 13
1 Electronic Records in Health Care 16
1.1 Aims of this chapter: 16
1.2 Introduction 16
1.3 The General Practitioner and the use of Electronic Medical Records 17
1.3.1 Global use of EMRs at GP Practices 17
1.4 Electronic medical records in the United Kingdom 26
1.4.1 Progress of EMRs in the United Kingdom 26
1.4.2 Health Technology Infrastructure 27
1.4.3 Major GP Software Systems used in the United Kingdom 28
1.4.4 Patients' views on using EMRs for healthcare 29
1.5 Description of the EMR Coding Structure in the UK 30
1.6 Explanation of the Quality and Outcomes Framework 31
1.7 Description of health research databases used around the World 33
1.8 Description of health research databases used in the UK 34
1.8.1 The Health Improvement Network database (THIN) 35
1.8.2 The Clinical Practice Research Datalink (CPRD) 35
1.8.3 The QResearch database (QResearch) 37
1.8.4 Research One 37
1.8.5 Linkages contained in the major UK databases 38
1.9 Advantages and Disadvantages of using EMRs for Research 39
1.9.1 Advantages of using EMRs in research 40
1.9.2 Limitations of using EMRs in research 40
1.10 Aim and Objectives 41
1.10.1 Study Aims 41
1.10.2 Study Objectives 42
1.11 The Gedling Cohort Electronic Medical Records 42
1.11.1 Introduction 42
1.11.2 Description of Gedling Cohort 43
1.12 Rationale of using EMRs in the Large and Small Databases in the Current Study 43
1.13 Overview of Next Chapters 45
1.14 Layout of Study 45

1.15 Next Chapter 46
2 Extraction from Primary Care Electronic Medical Records Collected from the Gedling Cohort 48
2.1 Introduction 48
2.2 Background to Gedling Study Population 49
2.3 Gedling Cohort Follow-up in 2000 50
2.4 Gedling Cohort Follow-up in 2007 50
2.5 Objectives of Using the Electronic Medical Records Collected from GP Practices 51
2.6 The Gedling Population Participating in EMR Data Extraction for the Current Study 51
2.7 The GP Systems Found at Nottingham Practices 53
2.8 Rationale for Choice of File Export Format of EMRs 55
2.8.1 The Export Options Available in the EMR Systems Encountered at GP Practices 55
2.9 The Potential Approaches to Processing Exported Electronic Medical Records and Extracting Data 57
2.10 The Selected Data Extraction Mechanism: Regular Expressions 60
2.11 Strengths and Limitations of Extraction Method 61
2.12 Previous Studies that have Used Regular Expressions to Extract from Electronic Medical Records 61
2.12.1 Formulating an Extraction Methodology Using Regular Expressions and Manual Review and Setting a Gold Standard 62
2.12.2 Accuracy and Completeness and Establishing the Gold Standard for this Study's Extraction Methodology 63
2.13 Explanation of Basic Regular Expression Notation 64
2.14 Programming Languages Available for Writing Regular Expressions 66
2.15 The Microsoft C# Programming Environment 67
2.16 Summary 68
3 Overview of the Conversion Process of Gedling Cohort Data Sets into a New Database 70
3.1 Introduction to the Extraction Software 70
3.2 Steps Users Take to Utilise Extraction Software 71
3.3 The Data Extract Stages 73
3.4 The role of primary and secondary regexes to do data extractions 74
3.4.1 Applying Primary and Secondary Regular Expressions 74
3.5 Rationale for the extracted data files structure 76
3.6 Extracting information about demographics 77
3.7 Extracting information on medical events 79
3.7.1 Extracting free text in the electronic medical records 80
3.8 Extracting information about drug prescriptions 81
3.9 Extension of dosages study 83

3.10 Uniquely identifying a patient 84
3.11 The Gedling Database 84
3.12 Summary 85
4 The Extraction Software Module for Each Electronic Medical Record Type and Assimilation of the Gedling Database 86
4.1 Aims of this chapter 86
4.2 Background 86
4.3 Producing the systmOne data file 88
4.3.1 Introduction to systmOne data extraction 88
4.3.2 Exporting records from systmOne 88
4.3.3 Problems encountered when extracting data 88
4.3.4 Data format and layout issues 88
4.3.5 Software programming issues 89
4.3.6 Splitting the systmOne file into subsections 90
4.3.7 systmOne demographic file 96
4.3.8 Extraction of Medical file data for systmOne 118
4.3.9 Extraction of Drug Prescriptions for systmOne 121
4.4 Producing the iSOFT Data Files 124
4.4.1 Introduction to iSOFT data extraction 124
4.4.2 Downloading the iSOFT files 124
4.4.3 PDF-text file conversion 124
4.4.4 Choice of PDF Conversion Tool 126
4.4.5 Problems encountered when extracting data 128
4.4.6 The iSOFT Demographics Extraction 130
4.4.7 The iSOFT Medical Extraction 140
4.4.8 Checking of Read or Clinical Codes 140
4.4.9 The iSOFT Drug Extraction 143
4.5 Producing the Emis LV Data Files 145
4.5.1 Downloading the Emis LV files 145
4.5.2 Problems encountered when extracting data 145
4.5.3 Splitting the Emis LV file into subsections 147
4.6.1 The Emis LV Demographic Extraction 150
4.6.2 The Emis LV Medical Extraction 171
4.6.3 The Emis LV Drug Extraction 174
4.7 Producing the Emis PCS Data Files 176
4.7.1 Introduction to Emis PCS data extraction 176
4.7.2 Problems encountered when extracting data 176
4.7.3 Downloading the Emis PCS files 176
4.7.4 Splitting the Emis PCS file into subsections 176
4.7.5 Demographics 177

4.7.6 Medical 196
4.7.7 Checking of Read or Clinical Codes 197
4.7.8 Drugs EMR section for Emis PCS 198
4.8 Producing the Emis Web Data Files 200
4.8.1 Problems encountered when extracting data 200
4.8.2 Downloading the Emis Web files 200
4.8.3 Splitting the Emis Web file into subsections 200
4.8.4 Marking the Boundaries of an Emis Web Record 201
4.8.5 The Main Emis Web Files Extracted 201
4.8.6 Demographics 203
4.8.7 Medical 216
4.8.8 Drugs 219
4.9 Other variations noticed between the different GP software systems 221
4.10 Post extraction: how the files were put together from the different GP systems 223
4.11 Strengths and Limitations 224
4.12 The Next Steps: using the EMR derived files in further analysis 225
4.13 Summary 226
5 Comparison of the Extracted Data in the Gedling Database across the different GP Software Systems 227
5.1 Aims of this Chapter 227
5.2 Introduction 227
5.3 Background 228
5.3.1 Measuring Key Characteristics of EMRs for Comparison 229
5.3 Methods 230
5.3.1 Study Population 230
5.3.2 Extracting age, sex, socioeconomic status and other factors 230
5.3.3 Computerisation Date 231
5.4 Results 235
5.4.1 Post computerisation adjustments for freetext, prescriptions and consultations 235
5.4.2 Free text 235
5.4.3 Prescriptions 239
5.4.4 Consultations 242
5.5 Discussion 246
5.5.1 Principal Findings 246
5.6 Comparison with other studies 247
5.6.1 Free text 247
5.6.2 Prescriptions 249
5.6.3 Consultations 251

5.6.4 Differences in median values for prescriptions and consultations.....252
5.6.5 The relationship between free text, prescriptions and consultations in the different GP systems.....253
5.6.6 Implications for using EHRS from GP systems for research.....255
5.7 Strengths and Limitations.....256
5.8 Summary.....257
6 Comparison of Extracted Medical Recordings of the Gedling Database to Cohort Questionnaire Follow-ups 1991, 2000 and 2007.....259
6.1 Introduction.....259
6.2 Background.....259
6.3 Data Collected from Previous Follow-ups.....261
6.3.1 Review of Gedling Survey Measurements 1991-2007.....262
6.4 Asthma and Chronic Obstructive Pulmonary Disease.....263
6.5 Rationale.....263
6.6 Methods.....263
6.6.1 Dataset & Study Population.....263
6.6.2 Defining the outcome code lists to examine Gedling cohort GP recordings 263
6.6.3 Statistical Methods and Analyses.....264
6.7 Results.....268
6.7.1 Self-reported asthma compared to asthma diagnoses in electronic medical records 1991, 2000 and 2007.....268
6.7.2 Asthma drugs recorded in electronic medical records compared to doctor-reported asthma at successive Gedling cohort surveys 1991, 2000 and 2007.....270
6.7.3 Diagnoses of asthma with a recording of an asthma drug in electronic medical records compared to doctor-reported asthma at successive Gedling cohort surveys 1991, 2000 and 2007.....270
6.7.4 Cases found for COPD from measurements collected in 1991 and 2000 compared to COPD diagnoses found in electronic medical records.....273
6.8 Discussion.....275
6.8.1 Principal Findings.....275
6.8.2 Strengths and Limitations.....275
6.8.3 Findings in Context of Previous Research.....277
6.8.4 Implications of Findings.....279
6.8.5 Summary.....282
7 Conclusions and Recommendations.....284
7.1 Summary of Main Findings.....284
7.1.1 Extraction from Primary Care Electronic Medical Records Collected from the Gedling Cohort into the Gedling Database.....284
7.1.2 Comparison of the Extracted Data in the Gedling Database across the different GP Software Systems.....285
7.1.3 Comparison of Extracted Medical Recordings of the Gedling Database to Cohort Questionnaire Follow-ups 1991-2007.....286

7.2 Main Implications of the Study 286
7.2.1 Extracting data from EMRs collected at GP practices 286
7.2.2 The effect of GP software on EMR recordings at practices 287
7.3 Recommendations for further studies 288
7.3.1 Large Scale Data Extraction 288
7.3.2 Exploring Free text in EMRs 289
7.3.3 Using EMR recordings to assess patient care 290
7.3.4 Encountering other primary care systems not covered 291
7.3.5 Combining clinical studies with GP data in future research 292
7.4 Concluding Remark on Fulfilment of Aims 293
7.5 Overall Summary 294
8 Appendices 295
8.1 Ethical approval for QOF/COPD study 295
8.2 Respiratory questionnaire 2000 296
8.3 Respiratory questionnaire 2007 311
8.4 Ethical approval for the Gedling Cohort Study 322
8.5 Letter of access to primary care practices 323
8.6 Sample letter to Practices 326
8.7 Extraction software on compact disc attached 326
8.8 Form and Dosage Lists 327
8.9 Downloading EMR steps 335
8.10 TRUD Dictionary: Both Read version 2 and Clinical Terms Version 3 (CTV3) ..336
8.11 Clinical Codes and Asthma Drugs 337
Read codes for asthma, version 3 339
9 References 346

List of Tables

Table 1 – Quality and Outcomes Framework points by domain in 2009/10, England (number of practices =8,305) with 2008/09 comparison 32

Table 2 - UK Health Databases and the potential linkages available for researchers 38

Table 3 GP Software System by Practice Visited During 2011-2012 55

Table 4 - Export Options for Electronic Patient Records (EMRs) at GP Practices Encountered During Visits to Collect Patient Records 56

Table 5 list of regular expression meta-character notation in Microsoft's C# Programming Environment..... 66

Table 6 - Full listing of patient table fields, character type and description of patient fields 78

Table 7 - Full listing of medical table fields, character type and description of medical-related fields 80

Table 8 - Full listing of drugs table fields, character type and description of drugs-related fields and those fields created by dosage extender software to match THIN dosage fields 83

Table 9 free text per person per consultation by vendor, age, IMD and social class.....237

Table 10 free text per person per consultation by vendor, age, IMD and social class post computerisation 238

Table 11 total prescriptions per person per practice by vendor, age, IMD and social class: 1991-Collection Date 240

Table 12 prescriptions per person per year per practice by vendor, age, IMD and social class: post computerisation 241

Table 13 total consultations per person per practice by vendor, age, IMD and social class: 1991-Collection Date 244

Table 14 consultations per person per year per practice by vendor, age, IMD and social class: post computerisation 245

Table 15 Data available from previous Gedling Follow-ups 1991, 2000 and 2007 and data available in Gedling cohort electronic medical records collected.....262

Table 16 diagnoses of asthma in electronic medical records compared to Doctor diagnosed asthma at successive Gedling cohort surveys 1991, 2000 and 2007 269

Table 17 Asthma drug recordings in electronic medical records compared to doctor-reported asthma in questionnaire results 1991, 2000 and 2007 271

Table 18 - Diagnoses of asthma with a recording of an asthma drug in electronic patient records versus doctor-reported asthma at successive Gedling cohort surveys 1991, 2000 and 2007 272

Table 19 Diagnoses of Chronic Obstructive Pulmonary Disease (COPD) in electronic patient records compared to case finding of COPD using Forced Expiratory Volume 1 (FEV1) recorded at successive Gedling cohort surveys 1991 and 2000 274

List of Figures

Figure 1 - Layout of how the current study achieves its objectives 46

Figure 2 a set of standard text with corresponding hidden characters exposed 64

Figure 3 target area for regular expression shown in grey highlight contrasted with rest of the standard text 65

Figure 4 A single regular expression targeted to locate the reference number of a customer in the fictional Penman and Howe limited 65

Figure 5 The different EMRs interacting with the extraction software to produce the three files of patient, medical and drug prescriptions 70

Figure 6 User interface of Extraction Software when running in computer memory and steps of use 71

Figure 7 Algorithm for extraction of EMR data used by the Extraction Software 73

Figure 8 - Regular expression highlights the higher-level sections in a whole record that then form the three key subsections 75

Figure 9 - secondary regex extraction from a subsection containing demographic values 75

Figure 10 - Structure of the Gedling Database Data Files linked by NHS Identifier 77

Figure 11 example of a complete systmOne record's overall structure 91

Figure 12 – systmOne: extraction of name, nhsid & date of birth 98

Figure 13 - systmOne: extraction of Gender, Place of Birth and Marital status 101

Figure 14 - systmone: extraction ethnicity, language and address 104

Figure 15 – systmOne: extraction of Registration Date 107

Figure 16 - systmOne: extraction of PDS Registered Practice 109

Figure 17 - systmOne: extraction of medical records, dispensing and pharmacy 112

Figure 18 - systmOne: final segment of record concerning the extraction of appointments, attendance, visits and did not attends 115

Figure 19 - systmOne: medical extraction date, Read code, Read code text and free text 118

Figure 20 – systmOne: drug extraction of date, drug name, dose, form, instruction and quantity 121

Figure 21 – iSOFT: extraction for Patient name, address and nhsId 131

Figure 22 – iSOFT: extraction of gender, date of birth, registration and date 136

Figure 23 – iSOFT: extraction of registered GP and postcode 139

Figure 24 – iSOFT: extraction of medical date, Read code, Read code text and any free text 141

Figure 25 – iSOFT: extraction of prescription date, drug name, dose, form and instruction 143

Figure 26 – Emis LV: Extraction of name, address & date of birth 152

Figure 27 – Emis LV: Extraction of number and NHS Id 155

Figure 28 – Emis LV: extraction of postcode and records at 157

Figure 29 – Emis LV: Extraction of hospital number, trading and registering Dr 160

Figure 30 – Emis LV: Extraction of usual dr, dispensing & RPP road miles 163

Figure 31 – Emis LV: Extraction of walking quarters, residential inst. & special blocked 166

Figure 32 – Emis LV: extraction of medical information of date, Read code text and any free text 171

Figure 33 – Emis LV: extraction of prescription date, name, dose, form and quantity ... 174

Figure 34 – Emis PCS: extraction of name, gender and address 179

Figure 35 – Emis PCS: extraction of date of birth postcode and marital status 182

Figure 36 – Emis PCS: extraction of nhsID, registered GP and usual GP 185

Figure 37 – Emis PCS: extraction of ethnic origin and trading partner 188

Figure 38 – Emis PCS: extraction of dispensing and residential instruction 190

Figure 39 – Emis PCS: extraction of hospital number, records at and walking quarters 193

Figure 40 – Emis PCS: medical extraction of date, Read code text, staff and Read code 196

Figure 41 – Emis PCS: drug extraction of date, drug comprising of name, dose and form 198

Figure 42 – Emis Web: extraction of patient name and date of birth 204

Figure 43 – Emis Web: extraction of nhsid and address 205

Figure 44 – Emis Web: extraction of registration date, usual GP and patient type 209

Figure 45 – Emis Web: extraction of language and dispensing 212

Figure 46 – Emis Web: extraction of advocacy and transport needs 214

Figure 47 – Emis Web: Extraction of medical subsection of date, Read code text and Read code free text 216

Figure 48 - Emis Web: extraction of prescriptions looking at date, quantity, instruction and drug information of name, dose and form 219

Figure 49 Extraction Software (ES) interaction with EMRs collected from GP practices and final rendering into 3 unified files covering demographics, medical and prescriptions extracted from all EMRs 223

Figure 50 determining the computerisation date of January 1st in relevant year by using asymptotic peak recording level as a proxy 233

Figure 51 GP System Adopted by Practice for those encountered during data collection from consented patients during the 2007 follow-up 234

List of abbreviations

Apollo Medical Systems (AMS)

Body Mass Index (BMI)

Byte Order Mark (BOM)

Cegedim Strategic Data (CSD)

Chronic Obstructive Airways Disease (COAD)

Chronic Obstructive Lung Disease (COLD)

Chronic Obstructive Pulmonary Disease (COPD)

Chronic Obstructive Pulmonary Disease (COPD)

Clinical Practice Research Datalink (CPRD)

Comma Separated Values (CSV)

Computer Sciences Corporation (CSC)

Database (DB)

Database Management System (DBMS)

Department of Health (DOH)

Did Not Attend (DNA)

DOB (Date of Birth)

Economic and Social Research Council (ESRC)

Electronic Health Database (EHD)

Electronic Health Record (EMR)

Electronic Medical Record (EMR)

Electronic medical records (EMRs)

EMIS (Egton Medical Information Systems)

Extensible Mark-up Language (XML)

Extraction software (ES)

Forced expiratory volume over 1 second (FEV1)

Forced vital capacity (FVC)

General Medical Services (GMS)

General Practice Research Database (GPRD)

General Practitioner (GP)

General Public Licence (GPL)

Health and Social Care Information Centre (HSCIC)

Health Data Warehouses (HDW)

In-Practice Systems (INSP)

International Classification of Disease (ICD)

Medical Research Council (MRC)

Medicines and Healthcare products Regulatory Agency (MHRA)

National Health Service (NHS)

Natural Language Processing (NLP)

NHS ID (National Health Service Identifier)

Oxford Medical Information System (Oxmis)

Personal Demographics Service (PDS)

Portable Document Format (PDF)

Postcode Variable Indicators (PVI)

Quality and Outcomes Framework (QOF)

Quality Management and Analysis System (QMAS)

Rural Practice Payment (RPP)

Structured Query Language (SQL)

Systemised Nomenclature of Medicine – Clinical Terms: (SNOMED-CT)

The Clinical Practice Research Datalink (CPRD) 37

The Health Improvement Network (THIN)

The Health Improvement Network database (THIN) 37

The QResearch database (QResearch) 38

Value-Added Medical Products (VAMP)

World Health Organisation (WHO)

1 Electronic Records in Health Care

1.1 Aims of this chapter:

- Introduce electronic medical records (EMRs)
- Introduce electronic medical records in the United Kingdom (U.K)
- Describe the systems used in UK primary care for EMRs
- Describe the Read coding scheme
- Introduce the Quality and Outcomes Framework (QOF)
- Describe health research databases used in the UK
- State the study aims and objectives

1.2 Introduction

Electronic Medical Records (EMRs) represent an important source of potential data useful for epidemiological research throughout the world. The longitudinal nature of their recording mean it's possible to assess how well care is being delivered to patients at the GP practice level. Within the UK, the introduction of the incentivised pay for performance Quality and Outcomes Framework (QOF) has had a major impact on GP practices both in terms of accelerated EMR systems adoption and priorities of care delivery. It is possible to use health databases like The Health Improvement network (THIN) to see how GP practices are fulfilling such QOF targets. There is also the potential to use an existing cohort of patients to explore what is happening in a locally built EMR-derived database. The major aim of this thesis is to assess the extent to which data is transferable from these different EMR systems and how well recording is consistent at GP level for QOF-covered disorders using both a large database and a purpose built locally created database.

1.3 The General Practitioner and the use of Electronic Medical Records

The role of the general practitioner (GP) in diagnosis is one of problem recognition and decision-making, this diagnostic process can sometimes act as a gateway to further management of the patient's concern for example referral to a hospital (1). Patient care is now being increasingly recorded electronically at GP practices (2). A typical primary care record takes the form of a dated history of a patient's interactions with their GP and other National Health Service (NHS) care and will contain their current and past prescribing records, diagnoses, lab test results, reviews, appointments, referrals and important demographic details, progress notes, problems, medications, vital signs, immunisations, lab data and radiology reports (3). As a consequence, the primary care medical record constitutes a near-complete medical history of each patient. Early years of physician documentation took the form of written narratives representing observations later evolving into individual paper records (4). By 2005 close to a 100% of all UK primary care settings used EMRs for recording patient care (5).

1.3.1 Global use of EMRs at GP Practices

The United Kingdom was one of the early adopters of EMRs compared to similarly developed countries such as other European nations, the US and Canada (6). The following introduction looks at EMRs around the world and for convenience purposes these are covered in the following regions:

Africa

- South Africa
- Kenya
- Uganda
- Ghana
- Tanzania

Europe

- France
- France
- Germany
- United Kingdom
- Italy
- Russia
- Sweden
- Denmark
- Holland
- Spain
- Norway
- Iceland
- Portugal
- Finland
- Lithuania
- Latvia
- Greece

Middle East

- United Arab Emirates
- Saudi Arabia
- Israel
- Egypt
- Turkey

South Asia

- India
- Pakistan
- Sri Lanka
- Thailand

Asia Pacific and South Pacific

- China
- South Korea
- Japan
- Australia
- New Zealand
- Indonesia

North America

- Canada
- United States
- Mexico

South America

- Argentina
- Brazil
- Peru
- Paraguay

1.3.1.1 Africa

Despite economic successes on the African continent South Africa is still a developing country made up of vast rural areas. In urban based health care computer technology is widely available whereas most rural areas do not having any computers at all (7). Such a distinctive digital divide is also apparent in other African based studies (8,9). One of these studies demonstrated EMRs being used in Kenya for Human Immunodeficiency Virus (HIV) care. Initially by creating a Microsoft Access spreadsheets linked by patient identifiers in 2001 but this became quickly overwhelmed by mid-2005 where 15,000 patients made up 50,000 visits. Its success meant the EMR implementation was then re-written in 2005 to become a freely available EMR platform. This was funded by the Rockefeller Foundation and the World Health Organisation (WHO) to test its usefulness to HIV and Acquired Immunodeficiency Syndrome (AIDs) programs in other developing countries of Tanzania and Uganda where it was successfully

installed and used at six sites (8). The second study showed that that the availability of EMRs in sub-Saharan Africa has been mostly driven by international efforts at stemming the HIV/AIDS epidemic (9). A recent review of EMR implementations in sub-Saharan Africa concluded that despite low resources the future is promising for EMR implementation and the use of technology to record health (10). There are no specific estimates available on EMR adoption rates across the African continent but it appears as if the general trend is towards EMRs being used mainly in the setting of laboratory, pathology and pharmacy information systems. For example the WHO Atlas survey 2015 estimated these types of EMRs to be in the region of 35%, 18% and 33% for Kenya (11). Similar percentages were estimated for South Africa (12), Uganda (13) and Ghana (14) respectively.

1.3.1.2 Europe

A 2015 study looking at the use of health care technology in 31 European countries reported that in 1993 results showed 40% of GPs surveyed in European countries reported using personal computers in practices. This is contrasted by the same study reporting that by 2012 almost all GPs surveyed used a computer (at 96%) but showing a variation in the general use of computers at practices (15). North and Western European countries were ahead in using EMRs in primary care and out of 31 European countries studied only five showed adoption rates below the 90% mark (15). The same study also showed that individual countries have different usage of computers at the GP practice level across Europe. For example in 2012, European GPs mainly used their computers to search for medical information on the Internet and issue prescriptions whereas in 1993 they mainly used computers for administration purposes only.

1.3.1.2.1 Variation in EMR use for appointments

The same study found another area of variation was using EMRs to in making appointments. It showed in 2012, 51% of European GPs used computers to do this function but is masked by variation as in Greece it is 4.5% and 100% in Holland for GP practices.

1.3.1.2.2 Variation in EMR use for consultations

Again another variation found in the same study was in the area of using computers for recording where in 2012 around 76% of GP practices in 31 European countries electronically stored consultations and stored diagnostic test results. The highest EU countries for storing both types of data were Holland, Denmark Spain, Norway, Iceland and the United Kingdom scoring 100%; followed by Portugal on 98% and Finland on 97% respectively. The lowest percentages were for Lithuania (19%), Latvia (24.5%) and 27% of Greek practices recorded both electronically. A different study reported that by 2004 90% of Sweden's GPs were using EMRs and in neighbouring Denmark that figure was 62% of GPs (16). Russia was a slower adopter of EMRs and is still seen as an emerging market for EMR adoption to an extent that is seeing many GP software companies entering to cater for an increasingly electronic patient care model in primary care. There is no specific adoption rate available but such is the interest in the Russian EMR market that in 2011 International Business Machines (IBM) became the service provider for nine Russian hospitals using its Lotus Notes software (17). In 2014, Moscow's Regional Government launched an EMR system through its United Medical Information and Analytical System of Moscow (EMIAS, in Russian) to automate booking of hospital visits, online appointments service and electronic prescribing and around 6 million people are covered in the system (17,18).

1.3.1.2.3 Patients' Access to their own EMRs

In the early 1990's Sweden was one of the earliest EU countries to give patients direct access to their EMRs in the Uppsala County Council area in 1997 (19).

Sweden's Uppsala County was also one of earliest examples of integration between primary and secondary care where all primary care centres and hospitals with the exception of community nurses and some private doctors used the same software system called Cosmic from a Swedish software company Cambio healthcare systems (19).

1.3.1.3 Middle East

Similar to Europe, there is variation in use of technology to deliver healthcare within the Middle East region. For example in the United Arab Emirates (UAE), Abu Dhabi's health authority has implemented an EMR system produced by Cerner Corporation in the capital Abu Dhabi and Al Ain in 2008 (20) . The use of EMRs was reported to be positive, however the major barriers to implementing EMR systems were identified as lack of computer literacy and protected time for training staff. In Egypt, the implementation of EMRs at practices in Alexandria found that post implementation EMR records were significantly less complete when compared to the paper-based records and the reasoning behind this suggested that high workload, complexity of the EMR systems and training were the reasons for discrepancies (21). Looking at the wider Arab world a study found that while 95% of physicians in Saudi Arabia reported using a Personal Digital Assistant (PDA) for work as a whole within the Arab region barely 1% of health services use this type of technology (22). The use of EMRs was high in Israel where it was reported that 98% of physicians use EMRs in public hospitals and that most Israeli hospitals use EMRs but there is no one particular standard (23). In Turkey, the Ministry of Health introduced The Health Transformation Program in 2003 which aimed to create EMRs for all citizens which would follow a set of standards and by November 2013, 98% of public hospitals and 80% of private and university hospitals were connected to the National Health Information System (NHIS) (24). The second release of Turkey's NHIS went further to integrate primary, secondary and tertiary care

which meant by the end of 2013 it contained more than 2 billion EMR documents and the EMRs of 78.9 million are contained within the NHIS (24).

1.3.1.4 South Asia

Data available on EMR adoption and use in South Asia was found to vary widely between countries even though they are geographically near to one another. There was no data available on Pakistan's use of EMRs though progress seems to be happening because the World Health Organisation (WHO) reports that Pakistan has adopted a national EMR strategy in 2013 and estimates that about 25-50% of primary care facilities use EMRs (25). However no further data was found or examples located for Pakistan. For Afghanistan there was again no data available on EMR use and according to a WHO survey this question was left unanswered so no estimates on EMR at primary, secondary or tertiary levels are available (26). In India the Federal Government indicated on its national website that there are plans to introduce a uniform system for using EMRs in hospitals and other health care providers though a specific time frame is not stated (27). A study looking at rural practices in India found that they generate a lot of data that would be stored in an EMR such as demographic data, vaccines monitoring, antenatal care which takes the form of reports but much of it is redundant and never utilized adequately (28). In India, primary care patient records are traditionally paper based with most providers still using this type of system with exception of secondary and tertiary care which use electronic records being highly dichotomous due to economic disparities (29). There were no specific estimates available for EMR adoption rates across India.

In Sri Lanka a recent study stated that though there is not a national EMR structure there is evidence that portions of medical records are being computerised especially in Government-led hospitals (30). There were no specific estimates for EMR adoption rates available for Sri Lanka as a whole.

In Thailand, its National Government made efforts in using low cost software resulting in the production of Hospital OS which is an EMR system implemented in 95 small rural hospitals and 402 health centres serving at least 5 million patients (31).

1.3.1.5 Asia Pacific and South Pacific

A recent China based study concluded that while there are few assessments on the level of EMR adoption across the Chinese health sector around 30% of hospitals use paper based records and those using EMRs do so using a diverse range of commercial products and that adoption of EMRs depends on hospitals' own income (32). In South Korea there has been a steep rise in the use of EMRs at Korean hospitals where by 2010 EMR adoption had reached 77% of hospitals. Japan has lagged behind compared to EMR adoption rates in Western Europe as a recent study indicated this level to be around 10% of GPs and hospitals using EMRs and where technology is utilised its mainly for billing purposes (33). One of the most populous Asia Pacific countries is Indonesia, where it is unclear on the proportion of EMR adoption nationwide but a national EMR policy does exist and was adopted in 2002 and has been partly implemented (34). In contrast the South Pacific region of Australasia; Australia and New Zealand have similarly high adoption rates for EMRs comparable with those found in Western and Scandinavian Europe. For example by 2006 Australia had EMR adoption rate of 98% for prescriptions and 64% exclusively writing clinical notes electronically and in New Zealand this figure was 90% for electronic prescriptions and 74% for electronic clinical notes with 99% of all New Zealand GP's using some form of GP software for administrative processes (6).

1.3.1.6 North America

The US Federal Government has encouraged EMR adoption under the Health Information Technology for Economic and Clinical Health (HITECH) Act 2009 resulting in a large number of US hospitals adopting these systems to

demonstrate meaningful use (35). There is evidence that these national policies are making a difference to EMR adoption in the US. For example in 2013, 59% of US hospitals had an EMR system (36) compared to a 2006 study that estimated EMR use in U.S. hospitals likely to be approximately 10% although a precise percentage was not available (37). In Canada, since 2006 EMR adoption rates have increased from around 20% of GPs to an estimate of 62% of GPs in 2013, with regional disparities ranging from about 40% of GPs in New Brunswick and Quebec to more than 75% of GPs in Alberta using EMRs (38). Both the United States and Canada still lag behind countries with near-universal adoption of EMRs. Although the spread of EMR use has been rapid in both countries, with a reported 50 percent increase in the use of EMRs from 2009 to 2012 (39). In Mexico, its Colima Region was one of the first areas to start using an EMR system called Management of the Electronic Health Record in Colima (SAECCOL, in Spanish) in 2005 (40). There are no official figures for EMR adoption across Mexico but one study reported that once the Colima region started using EMRs, neighbouring state of Coahila also adopted the same system in their hospitals with other Mexican states expressing an interest at the time (41).

1.3.1.7 South America

There are no specific data on EMR adoption rates in Argentina however a study did consider an example where an EMR system in the Hospital Italiano of Buenos Aires (HIBA) which was used to enter problems and found a higher level of accuracy using the EMRs compared to other settings in the same hospital (42). Again for Brazil there is no specific set of studies specifying EMR adoption figures. However Brazil has implemented urban electronic health care on a large scale. For example In Sao Paulo, the Saúde Health Information System serves 14 million registered patient and stores data on 20 million patient encounters from 702 health facilities yearly (31). In neighbouring Peru, there are no specific data available on EMR adoption however a WHO survey does indicate that secondary and tertiary care facilities have an EMR usage of around <25%

(43). In nearby Paraguay the situation seems a lot better in comparison. A WHO survey indicates that Paraguay adopted a national EMR infrastructure in 2010 and estimates that primary care facilities using EMRs to be in the region of 50-75% and secondary care facilities being at 25-50% with tertiary care facilities using EMRs to be at <25% (44).

1.4 Electronic medical records in the United Kingdom

1.4.1 Progress of EMRs in the United Kingdom

The next section looks at the development of EMRs in the United Kingdom in more detail. EMRs have a long history of use in the UK. Especially since the 1970's as a number of individual NHS Trusts and hospitals introduced their own information systems having evolved from earlier systems created to manage and support administrative functions; such as maintenance of the patient register (and storage of basic demographic and residential details of patients) and appointment booking (45). Over time, these systems expanded to store increasing volumes of coded and narrative information about patients, their health, symptoms and diagnoses, the treatments prescribed and the results of tests and measurements ordered (46). In 2008 the level of EMR adoption was reported at 89% across the U.K (6) and in 2014 the level of EMRs adopted at practices in England is near universal at 99% of English GP practices (15). EMR Adoption has been driven by a number of factors: promotions of free computers, availability of inexpensive software, government funding of capital and recurrent EMR costs (6). The most recent 2004 National Health Service (NHS) contract with GPs created a set of quality measurements and improvement requirements as part of remuneration meaning that computerisation of patient records was further incentivised (47). Ensuring compliance to these measures is an important area of focus for GP systems. For example in one practice a GP software called Vision has a mechanism of prompts to remind GPs to ensure recordings are done to maximise compliance to incentivised targets (48) and this

is also used as a selling point with a popular GP software called systmOne (49). There is evidence of incentives also encouraging interoperable working between GPs at the primary care level and hospitals at the secondary care level. For example when a patient is discharged from hospital a discharge summary is made available to the patient's GP within 48 hours or the hospital incurs a financial penalty at the end of its fiscal year (47). Another important incentivised scheme is called General Practitioner System of Choice (GPSoC) which is a is a contractual framework through which GPs can procure IT systems/services from accredited suppliers and was launched by the Health and Social Care Information Centre (HSCIC) on behalf of NHS England (50). The scheme provides a choice of centrally and locally-funded IT systems and services presented in an online catalogue and opportunity to hold GP system suppliers to account for the delivery of service improvement plans and funded training for national services such as the Electronic Prescription Service (EPS) and the Summary Care Record (SCR) (51).

1.4.2 Health Technology Infrastructure

The National Programme for IT in the NHS (NPfIT) launched in 2002 was the largest public IT programme ever attempted in the UK with the core aim of providing healthcare practitioners with electronic access to patients' records. Two specific EMR applications termed the Summary Care Record (SCR) and the Detailed Care Record (DCR) were introduced in order to support electronic record keeping and sharing. The SCR was to allow healthcare providers in any part of the country to access electronic information on patient demographics and essential healthcare (e.g., allergies and current medication) (52). However, the implementation was marked by delays, implementation problems and stakeholder opposition with the programme finally being dismantled in 2011 with the Government Major Projects Authority (MPA) concluding that NPfIT was not fit for purpose (52). The Health and Social Care Information Centre was set up as

an Executive Non-Departmental Public Body in April 2013 under The Health and Social Care Act 2012 with responsibility to support the delivery of IT infrastructure, information systems and standards to ensure information flows efficiently and securely with the Summary Care Record , the Electronic Prescription Service and IT systems for GP surgeries and also Hospital IT systems (53).

1.4.3 Major GP Software Systems used in the United Kingdom

The major GP software suppliers for the UK are systmOne, Emis, Vision, iSOFT and smaller scale suppliers are Evolution and Cross care. A brief background on each one:

1.4.3.1 SystmOne

The Phoenix Partnership's (TPP) systmOne GP software has seen a large growth in their GP market share starting with around 20 practices in 2001 to covering 200 practices by 2006 and over 1000 practices by 2011 by which time they were the second largest GP supplier (54). TPP also now provides access to records on Android and Apple devices using its SystmOnline app (55).

1.4.3.2 Emis

This is a software supplier that originated in Egton, North Yorkshire in the 1980s and has since become a major player in the GP software market. Dr Peter Sowerby and Dr David Stables developed the software with the aim that it should improve patient care with the motto of Emis being software 'written by doctors, for doctors' (56). The group's first GP clinical software named Emis, had a character based user interface and was developed from 1987, with full commercial roll-out taking place in 1990 (57). Emis has three editions widely used in the UK: Emis LV, Emis PCS and Emis Web and as of March 2014 Emis LV and Emis PCS systems are being gradually withdrawn to focus solely on Emis Web as its flagship edition as Emis Health states that Emis LV and Emis PCS cannot be improved further for new GPSoC 2014 requirements (58).

1.4.3.3 Vision (In-Practice Systems, INSP)

In 1995, Reuters launched Vision, a new Windows-based practice management software application. In 1999, Reuters' practice management software business was acquired by Cegedim, a European healthcare software and research company, and renamed it In-Practice Systems (59). Vision has recently started to allow GPs to access patient records on mobile smartphones using the Vision Anywhere app (60).

1.4.3.4 ISOFT

ISOFT has been part of several acquisitions and mergers since its founding. Its flagship software package 'Lorenzo' has been a major part of the NHS National Programme for IT and in April 2011 iSOFT was acquired by Computer Sciences Corporation (61) and their healthcare software products are also being used in New Zealand practices and in some parts of the United States Military (62).

1.4.3.5 Evolution Practice Manager for Windows

Microtest Limited established in 1955 and in later years produced Evolution Practice Manager which has a graphical user interface for interaction between the GP and computer system and covering processes of recording, dispensing, journal entries, auto Read codes and free text notes (63).

1.4.3.6 Cross Care

By Advanced health and care limited. They provide two products for GPs it is Crosscare Primary and Adastra software which is aimed at secondary care providers. The Crosscare Primary system can also be customised for four editions covering hospice care, private care, primary care and corporate care (64).

1.4.4 Patients' views on using EMRs for healthcare

Since April 2014 GP practices were required to promote and offer the facility for patients to view their own EMRs online or print the information held in their Summary Care Record (i.e. medications, allergies and adverse reactions and any additional information the patient had explicitly consented to) and offer the

facility for patients to book, view, amend, cancel and print appointments online (65). One study that looked at patient's perceptions around EMRs found that for the U.K there are significant differences in levels of support depending on socio-demographic characteristics with older participants and those who are black British showing less support for the use of EMRs being nationally available and being used for multiple purposes such as personal health care, policy planning and health research but overall there is high support for the use of EMRs for personal health care (66).

1.5 Description of the EMR Coding Structure in the UK

Computer systems that are used to record medical data at the GP level typically require information to be entered in the form of codes representing each disease, health event, or patient characteristic of interest. This helps to avoid confusion due to variations in terminology, and makes records easier to search, classify and manage. Consistent coding of disease is important for the management of health care and auditing of the quality of care provided (46). The global standard endorsed by the World Health Organisation (WHO) is referred to as the International Classification of Disease (ICD) (67). In the UK there have been four types of code used to record diagnoses: 'Oxmis' (Oxford Medical Information System), Read code, 'local codes' (generated by GP systems) (68) and Systemised Nomenclature of Medicine – Clinical Terms (SNOMED-CT). Oxmis codes preceded Read codes and were used until the late 1990s, and Read codes incorporate all ICD codes, plus a large number of additional terms selected as a consequence of their regular use in primary care (69). When recording details of a consultation, GPs must select one or more codes from a list of possibilities. In addition, many clinical computer systems automatically choose and insert codes based on numeric recordings, test results or measurements made on patients (which can be overridden or modified by the

practitioner if necessary) and also allow the entry of richer detail in a 'free text' or narrative form (46).

The incentivised remuneration scheme using care targets to determine GP pay defines a list of Read codes which should be (exclusively) used to record details of monitored conditions and procedures (70). As a result Read codes have become the de-facto standard for the UK, as compliance with care targets requires that practice management software must enable either direct entry of qualifying Read codes, or provide 1 to 1 equivalents within their own coding schema. For historical reasons, older standards (e.g. Oxmis codes) and custom coding schema (local codes) persist; for example, the Emis EMR package provides a number of "Egton codes" used to record certain types of non-QOF data. Therefore, in extracting any set of diagnoses, it is necessary to capture all possible relevant codes consistent with the diagnosis and manually determine which are relevant (71). It is expected that SNOMED-CT coding system will eventually replace the current 'Read code' structure as the UK standard (72). This may be beneficial for international comparisons as research has shown there is a greater degree of direct mapping between SNOMED-CT and the ICD system (68).

1.6 Explanation of the Quality and Outcomes Framework

The quality of care and patient management in UK primary care medical practice is assessed according to the Quality and Outcomes Framework (QOF). QOF is an annual reward and incentive programme for all primary care providers in the UK introduced as part of the General Medical Services (GMS) contract for General Practitioners on 1st April 2004. Whilst the GMS contract does not oblige practices to participate in the QOF pay-for-performance scheme, uptake is high due to the large payments available; in 2009/2010 English practices received an

average of £126.77 for each of 1,000 "points" awarded against specific outcome measures (73). 8,305 practices (covering almost 100% of the registered patient population of England) were included in the published results. On average, practices received payment for 94% of the available points (74).

Table 1 – Quality and Outcomes Framework points by domain in 2009/10, England (number of practices =8,305) with 2008/09 comparison

QOF Domain	Average Points per Practice 2009/10	Average % Points Achieved 2009/10	Average % Points Achieved 2008/09
Clinical	668.2	95.9% (of 697)	97.8% (of 650)
Organisational	161.4	96.3% (of 167.5)	95.8% (of 167.5)
Patient Experience	65.4	71.5% (of 91.5)	84.2% (of 146.5)
Additional Services	42.0	95.3% (of 44)	97.3% (of 36)
Total QOF	**936.9**	**93.7% (of 1,000)**	**95.4% (of 1,000)**

Table ©2010 HSCIC and reproduced from http://www.hscic.gov.uk/catalogue/PUB04431/qof-09-10-rep.pdf (73) *in compliance with Re-Use of Public Sector Information Regulations (non-commercial reproduction).*

Achievement data for QOF 2015-16 shows that average achievement score for practices was 532.9 points out of 559 with the highest levels of achievement were for obesity and chronic kidney disease where 99.9 per cent was achieved lowest level of achievement was for Osteoporosis at 87.5 per cent. In 2015-2016 a total of 640 practices achieved the maximum of 559 points. A large increase on previous QOF year of 2014-15 where 448 practices achieved the maximum of 559 points for that QOF reporting year (75). QOF points are awarded in four performance domains (73):

- Clinical Domain: 86 indicators in 20 areas (Asthma; Atrial Fibrillation; Cancer; Cardiovascular Disease - Primary Prevention; Chronic Kidney Disease; Chronic Obstructive Pulmonary Disease; Coronary Heart Disease; Dementia; Depression; Diabetes; Epilepsy; Heart Failure; Hypertension; Hypothyroidism; Learning Disabilities; Mental Health;

Obesity; Palliative Care; Smoking; Stroke and Transient Ischaemic Attacks) worth up to a maximum of 697 points (69.7% of the total).

- Organisational Domain: 36 indicators in 5 areas (Records and Information about Patients; Information for Patients; Education and Training; Clinical and Practice Management and Medicines Management) worth up to a maximum of 167.5 points (16.8% of the total).
- Patient Experience Domain: 3 indicators in 2 areas (Length of consultations, Patient surveys and Patient experience of access) worth up to a maximum of 91.5 points (9.2% of the total).
- Additional Services Domain: 9 indicators in 4 areas (Cervical Screening; Child Health Surveillance; Maternity Services and Contraceptive Services) worth up to a maximum of 44 points (4.4% of the total).

For each condition in the clinical domain, practices are required to maintain a register of affected patients, and to report their compliance with a range of performance targets (the "indicators") among these patients. The complexity of the targets and of the data collection process is such that it has effectively become mandatory for primary care practitioners to maintain their patient records in electronic form using a software package capable of generating an accurate QOF return.

1.7 Description of health research databases used around the World

The rapid adoption of EMRs across different regions of the World means there was also was a corresponding increase in the number of countries where organisations have been able to link these or otherwise extract EMR data in a manner that has produced large databases which are rich in patient or participant related data. Although this study looks using a U.K database there are other similar research databases that exist around the World (76) including:

- Veterans Affairs claims data

- US Renal Data System
- Studies of Left Ventricular Dysfunction (SOLVD)
- Scandinavian cancer registries
    - The Danish Cancer Registry
    - The Cancer Registry of Norway
    - Swedish Cancer Registry
- Japan Medical Data Centre (JMDC)
- Saskatchewan Health Canada

The above health databases have different uses and the main areas of interest are (76) include:

- Burden (cost) of illness and cost of treatment
- Treatment patterns
- Medication adherence, persistence, and compliance
- Comparative effectiveness and costs of alternative treatments
- Disease prevalence and public health impact
- Safety studies

1.8 Description of health research databases used in the UK

There are also primary and secondary care EMRs used in large databases in the United Kingdom which contain prescribing, medical and demographic details. In recent years, the widespread use of EMRs in the UK has been exploited to create large health research databases. Medical records stored by large numbers of general practices around the UK have been anonymised (by removing patient names, addresses and day of birth, with month and year still present) and aggregated to extract populations, which can be used to carry out large-scale, statistically powerful epidemiological studies without the need to contact individuals directly. These resources enable appropriately designed studies to be carried out far more cheaply and rapidly that would have hitherto been possible.

In the UK, there are four main databases created in this way: The Health Improvement Network (THIN), The Clinical Practice Research Datalink (CPRD), QResearch and Research One. For compatibility reasons, each is created using data from practices using a single brand of computer system. Data from the 'Vision' package is used in the CPRD and THIN databases and data from Emis forms the basis of the QResearch database (77) and data from systmOne EMRs makes up Research One (78). The linkage within these databases is discussed next.

1.8.1 The Health Improvement Network database (THIN)

THIN is a database of electronic medical records from over 1500 GPs in over 500 UK practices. Diagnoses are recorded using Read codes and prescriptions are recorded using codes from the Prescription Pricing Authority (79). As of 2014, THIN had over 11 million patients, of which almost 3 million are actively registered. The remaining patients have historical data but have either left the practice or died (80).

1.8.2 The Clinical Practice Research Datalink (CPRD)

The Clinical Practice Research Datalink (CPRD) was initiated in 1987 starting out as the Value-Added Medical Products (VAMP) database. In 1993 it was acquired by Reuters and donated to the UK Department of Health where it was then renamed the General Practice Research Database and, later, as the Clinical Practice Research Datalink. The CPRD is a widely used UK-based database of clinical primary care records, and has been extensively used in primary care research. It is one of the world's largest sources of anonymised longitudinal data from primary care (81,82) and currently contains information on 13.6 million patients and like all the databases mentioned these exist in a dynamic cohort where patients can join/leave over the course of follow-up (83). Data is available on over 5 million active patients from approximately 650 general practices in the UK (82). The data supplied by each practice undergoes quality control and validity checks by the Medicines and Healthcare products Regulatory

Agency (MHRA) before being made available to researchers. CPRD has resulted in 1,500 publications which have led to improvements in drug safety, best practice and clinical guidelines (84).

1.8.3 The QResearch database (QResearch)

The data currently comes from approximately 1000 general practices using the EMIS clinical computer system and includes data from patients who are currently registered with the practices as well as historical patients who may have died or left. Records extend back to the early 1990's making it one of the largest and extensive general practice databases in the world (85). It contains the EMRs of over 7.5 million patients in total with approximately 3 million registered patients with the others having left or died (86).

1.8.4 Research One

ResearchOne was unveiled in May 2012 and was the result of a partnership between the University of Leeds via the Knowledge Transfer Partnership scheme and the resulting database consists of de-identified clinical and administrative data drawn from the potentially 20 million plus electronic patient records currently held on the TPP systmOne clinical system (87). Six months after its official launch in February 2013 there were 5 million non-identifiable records approved for research (88). The research records contain linked data from a wide variety of settings across both primary and secondary care.

1.8.5 Linkages contained in the major UK databases

The main UK databases have linkages with other data variables outside of the main areas of only GP documented diagnoses and treatment. These extra data variables wider scope of research. These are main linkages for the above UK databases:

Table 2 - UK Health Databases and the potential linkages available for researchers

Health Database	Linkages to other data variables
THIN	• Pollution indices • Townsend deprivation index • Hospital episode statistics (HES) • Census indicators for rural urban residences (89) • Geochemical baseline survey of the environment (90) and Experian's Mosaic consumer segment classification (91). • The anonymised patient data are collected regularly from the practice's Vision clinical system and sent to Cegedim Strategic Data (CSD) who then supply the THIN Data to researchers for studies and also provide anonymous postcode linked area based indicators of deprivation, ethnic mix and environmental exposures (92).
CPRD	• Hospital episode statistics (HES) • Death registration data • Office of National Statistics (ONS) mortality data • Cancer registry cardiovascular disease register deprivation at the Lower Layer Super Output Area based on patient and practice postcode (93).
Health Database	**Linkages to other data variables**

Q Research	• The database is linked to cause of death data at individual patient level with linkages extending back as far as 1993 with approval given to link QResearch to cancer registries and hospital episode statistics (85). • Socio-economic details of each patient's postcode • Cause of death data ONS mortality • Cancer registration data • Ethnicity from 2001 census (94).
Research One	• Census and demographic indicators (95) • Index of multiple deprivation (at Lower Layer Super Output Area Level) and link to Hospital Episode Statistics (HES) is under development (95).

1.9 Advantages and Disadvantages of using EMRs for Research

The increasing adoption of electronic medical records (EMRs) offers the opportunity to permit rich data collection and longitudinal analysis of patients and to increase patient care and patient-oriented research activities; whilst also reducing some burden on physicians, patients, and healthcare delivery (96). On balance using EMRs for secondary clinical research is something which has huge potential for success and also some limitations. The use of EMRs for health research has both potential advantages and limitations and examples of these include:

1.9.1 Advantages of using EMRs in research

- Having patient data stored electronically increases the availability of data, which may lead to more quantitative analyses to identify evidence-based best practices more easily (97).
- May enable a better understanding of health patterns in relation to specific medications with potentially much larger and more diverse members of the public, and to understand which treatment offers the best results (98).
- Having electronic health information about the entire population of patients they can mean a more meaningful look at the needs of patients who suffer from a specific condition, are eligible for specific preventive measures or are currently taking specific medications (99).
- EMRs can be used to automate pre-screening and, in turn, can significantly improve recruitment efficiency and reduce costs by excluding ineligible patients who do not merit further review (96).

1.9.2 Limitations of using EMRs in research

- If research is based on specific EMRs it raises the question about whether or not many of the benefits identified can be generalised to other settings of care (97). For example this would mean that specific EMR results in U.K are not applicable to the United States.
- EMRs typically contain information obtained through routine data collection collected in the process of providing clinical care while to make valid research conclusions may require specialised data collection with more precision. For example, a routine blood pressure measurement recorded in the EMR in a hospital may not be suitable for inclusion in a clinical research trial dataset because the patient's body position (e.g., sitting, standing, spine) was not documented (100).
- The study also indicated that even with patient authorisation, when research staff external to the organisation need access to patient

information, they may not be able to obtain it for practice-based opposition (96).

1.10 Aim and Objectives

1.10.1 Study Aims

This study aims to use electronic health data in two different research projects. The initial aim of this thesis is to become familiar with the structure and format of research databases by carrying out an exemplar research study using an existing large database of primary care medical records (The Health Improvement Network database). The second aim is to obtain the EMRs of the Gedling cohort participants and to convert them into a standardised format suitable for use in future research projects. This was accomplished through a series of staged objectives.

Overall the main outcome is to develop instructions and software to enable researchers to build bespoke GP EMR databases for their own cohorts of patients by visiting patients' practices and extracting data directly from their EMR installation. Several different EMR software packages are used in the UK therefore it will be necessary to:

- Define a standardised EMR data format
- Develop software to convert extracted records into this format

This will be done in the following stages:

1. Becoming familiar with the data recording format of an existing, widely used, EMR research database (chapter 1)
2. By using an existing, Nottingham based Gedling cohort to:
    a. Identify the EMR software package in current use
    b. Determine how to extract data in the most readily processed format
    c. Document this process
    d. Describe the structure of the data extracts (chapter 2-3)

3. Defining a single data storage format (with a structure similar to an existing widely used research database) capable of holding data extracted from these different EMR packages and developing software to convert the data to this new format (chapter 4).
4. Describing the quantity and recording quality of the processed data in the overall cohort and broken down by the software package the records originated in (chapter 5).
5. Demonstrating the practical use of the processed data by carrying out a comparison between self-reported diagnoses and indicators of health and those recorded by GPs (chapter 6).

1.10.2 Study Objectives

- Use data from a General Practice database, The Health Improvement Network (THIN).
- To contact the practices at which the patients enrolled in the Gedling cohort are registered, and to obtain copies of their Electronic medical records (EMRs).
- To develop software platform to automate the conversion of extracted records into the new format.
- To evaluate the similarities and differences between records originating in different EMR software packages, and to define a common data format into which records from each may be readily converted, and which is as similar as possible to existing research databases (and therefore easily utilisable by researchers familiar with them).
- To compare the quantity and quality of data originating from the different EMR systems encountered.

1.11 The Gedling Cohort Electronic Medical Records

1.11.1 Introduction

An existing Nottingham-based cohort which has given consent for their EMRs to be collected and analysed will be used as the basis for the small scale health database so that EMR recordings can be analysed and comparisons made. A short description is provided next and a detailed account of the cohort is presented in chapter 3.

1.11.2 Description of Gedling Cohort

In 1991, a random sample of residents of the Gedling district of Nottingham were invited to take part in a study of the relation between diet and lung disease (101). The participants were followed up again in 1999 (102) and 2000 (103). In order to expand the range of research projects that could be carried out in this cohort, in 2009 participants were asked to give researchers permission to access to a copy of their GP record. As the participants in the Gedling cohort were selected at random from the electoral register, their medical records are held at a large number of primary care practices which use a range of different EMR systems. It will require visiting the practices and collecting the EMRs and analysing the records and extract data from those EMRs to produce the Gedling Database. Further and detailed background information is included in chapter 3.

1.12 Rationale of using EMRs in the Large and Small Databases in the Current Study

Different GP systems record data in their respective EMRs differently, in terms of layout, terminology, location of routine results and general user interface. The U.K Government has contributed to a standardised recording in EMRs containing incentivised conditions so they can carry out pay for performance assessments enabling GPs to be remunerated. Due to EMR differences specificity and sensitivity definition vary between different GP packages however when QOF's pay for performance was introduced the NHS indirectly helped to solve this by standardising way data about conditions of interest should be recorded. Therefore in a database such as the bespoke one that is being produced from

this study, data from mixed sources regarding QOF is complete. Conditions of interest covered in the pay for performance framework of QOF will be most comparable as its data requirements are common to all GP records. Therefore it is important to know whether patients recorded as having diseases according to QOF criteria really do represent the complete population of those who have the condition or whether some patient have evidence of the condition but are not included in the QOF register. If such patients exist it is also important to determine whether they differ substantially from those included in the QOF register in any important respect. This is especially important when using data from mixed sources as this group of patients will be the most vulnerable to differences in quantity and quality of data recorded that are caused by differences in the software packages used to record the records. It can be assumed that there is some non-QOF patients' data in different GP software systems but whether it is same quality is an area that can be considered in this study. Due to the popularity of different EMR packages can vary geographically there is a risk that if this group is large it can introduce bias into studies using EMR data. Using an existing large scale database of extracted EMRs and assessing how well it actually fulfils criteria of those areas which have been incentivised and comparing findings to data extracted from real GP records in a smaller database for problems covered by QOF and disorders which are non-QOF we can gauge an idea of the effect produced in GP records from the introduction of QOF. We can then consider the implications and usefulness of using EMRs in both large and small scale databases as a result and whether there strengths and limitations in either approach and whether this could be a model for producing useful results from GP records collected at practices. To enable use of collected EMRs this study will develop a software-based approach that permits extraction of data from the electronic patient records collected into working data tables, and to then compare the output.

1.13 Overview of Next Chapters

A description of the subsequent chapters:

Chapter 2: Background and introduction to the Gedling cohort history up to the present study and initial steps taken when visiting GP practices to download EMRs and options considered to extract data from the collected EMRs.

Chapter 3: Background and introduction to the GP software systems from which Gedling cohort EMRs were collected and issues surrounding data extraction and outline and justification for methods used to extract from Gedling cohort EMRs.

Chapter 4: Description of processes that occurred to enable data extraction from Gedling cohort EMRs by GP software system. Issues encountered and how the resulting data files were put together and a rationale and justification of their design.

Chapter 5: Comparison of EMR extracted data from the Gedling cohort across the different GP Software Systems.

Chapter 6: Comparison of EMR extracted data from the Gedling cohort to previous follow-ups from 1991-2007 and its questionnaire results to existing EMR recordings in the present study.

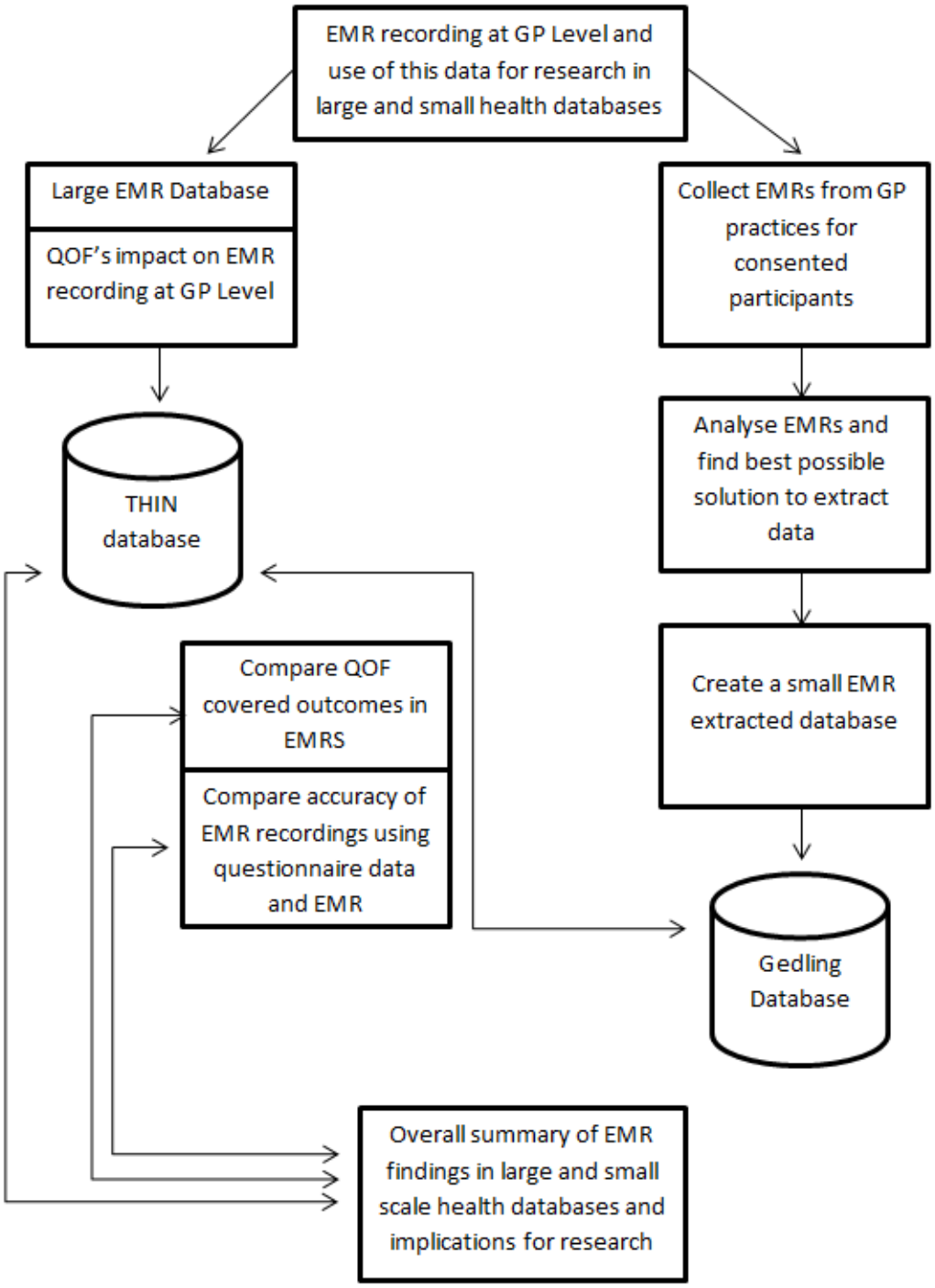
Chapter 7: Concluding remarks and summary of key findings from database studies using THIN and Gedling Cohort EMRs. What the implications are for epidemiological research using EMRs and recommendations for the future.

The above chapter descriptions are presented next as a flow chart.

1.14 Layout of Study

Flow chart of how the above chapters will fit together in Figure 1:

Figure 1 - Layout of how the current study achieves its objectives



1.15 Next Chapter

The next chapter looks at the impact of the U.K's incentivised health programme on EMR recordings by using an existing large health database widely used for

health research. It will also allow the results from the next chapter to be used when making comparisons to outcomes in the Gedling database that will be produced in chapters later on in this study.

2 Extraction from Primary Care Electronic Medical Records Collected from the Gedling Cohort

Aims of this chapter:

- Introduce and describe the Gedling study population
- Describe the experience gained from visiting GP practices
- Describe the EMRs software systems found at the practices
- State the export options available in the EMR systems encountered
- Describe the potential approaches available to extract data from the EMRs collected
- Justify the choice of extraction mechanism
- State the basic language notation of regular expressions used in the data extraction process
- Describe the programming tool used for development of the Extraction Software (ES)

2.1 Introduction

This chapter looks at extraction of data from the EMRs collected at various Nottinghamshire GP practices for the purposes of this study. A primary concern in building an extraction tool has been to create software that can be used by researchers at the University of Nottingham so that extracting medical-related events from EMRs becomes a gap that is filled rather than an expensive and time-consuming constraint on research that looks at cohort-related EMRs. The Gedling cohort is being used for the patient-consented EMRs and THIN as a template to create the new health database. THIN is being used as a template because it is widely used and researchers are familiar with its format and so their ability to transfer their knowledge and re-use existing programming scripts exists. Chapter 3 looks at what has happened during previous follow-ups of Gedling cohort, sets the objectives of using the collected EMRs, the experiences

of collecting the EMRs and examines options to extract EMR data and justifies the rationale behind the approach chosen.

2.2 Background to Gedling Study Population

The original 1991 study that set-up the Gedling cohort aimed to examine the relationship between diet and respiratory disease. The 1991 study population was identified by systematic sampling from a random starting point in the electoral register of Gedling, a local authority area in Nottingham. From a total population of approximately 87,000 adults, the initial study identified 7,106 names and wrote to each asking them to participate in the study. The final sample size was 2,633 adults aged 18–70 living in Nottinghamshire, U.K. A total of 4 papers mainly focusing on the link between diet and lung function were published from this 1991 cohort (101,104–106). The information collected from the 1991 sample included:

- Food frequency questionnaire
- Smoking exposure by personal recall
- Age
- Sex
- Height
- Postcode
- Forced expiratory volume over 1 second (FEV1)
- Forced vital capacity (FVC)
- Airway reactivity to methacholine
- 24hr urinary sodium excretion
- Skin prick tests to dermatophagoides pteronyssinus, grass pollen, cat fur and aspergillous and claudisporium
- Medical and occupational details
- Computerised questionnaire on respiratory symptoms

The questionnaire on respiratory symptoms is included in appendix 8.2. 1991 was questionnaire was virtually the same for 2000. The 2000 questionnaire is the closest available to the 1991 questions. There was different question on eczema being the main difference between 1991 and 2000. This study makes use of the data from the 1991 computerised questionnaire (2000 shown as example)

2.3 Gedling Cohort Follow-up in 2000

Nine years later all surviving individuals were invited to participate in a follow-up study. Non-responders were sent two reminders by post and, where possible, contacted by telephone. Of the 2,633 subjects who took part in the original survey in 1991; 1,346 (51%) provided measures in 2000. Of the 1,287 who did not participate in the 2000 study, 188 (15%) had moved outside the Nottingham area, 139 (11%) had died or were unable to attend because of chronic ill health, 475 (37%) replied that they were no longer interested in participating, and 95 (7%) agreed to participate but did not attend. Researchers were unable to establish contact with 398 (31%) subjects. Participation in 2000 was lower in the younger age groups. Males 49.60% (n=666), females made up 50.40% (n=677) of the sample. Two papers one focusing on the link between diet and lung function and a second on increasing prevalence of skin sensitivity were published from this 2000 cohort follow-up (107,108) respectively. The information collected from the 2000 sample was the same as in 1991 with the exception for airway reactivity to methacholine. The computerised questionnaire data from 2000 used in this study and is available in appendix 8.2

2.4 Gedling Cohort Follow-up in 2007

In 2007, surviving patients were again contacted and participated in a questionnaire based study investigating the role that dietary fatty acids may

play in airways disease. The information collected from the 2007 sample included:

- Food frequency questionnaire developed to also include information on fatty acids.
- Respiratory questionnaire.
- Permission to be sent a copy of their death certificate
- Permission to access a copy of their EMRs within the next 5 years period

Again in the same manner for 1991 and 2000, the 2007 computerised questionnaire was used in this study is available in appendix 8.3

2.5 Objectives of Using the Electronic Medical Records Collected from GP Practices

A large number of the 2007 cohort patients gave permission for their EMRs to be copied. EMRs are now primarily in electronic format and are stored in a variety of GP software systems. Once all consent forms were collated the next steps were to establish the objectives of the project and these were therefore to:

- Visit each participant's practice
- Identify software systems in use and available export formats
- Choose most complete and open format for each system
- Securely copy and transport the EMR
- Convert data to a standardised format suitable for research use

2.6 The Gedling Population Participating in EMR Data Extraction for the Current Study

- 1,346 patients provided measures in 2000. In the 2007 follow-up, 73% (n=978) of the participants gave permission at 140 practices nationwide to medical record collection.
- The author identified 30 practices covering 80% of consenting patients (n=781) which were in Nottingham postcodes and selected for inclusion

due to being geographically feasible to collect within a reasonable study time frame before expiry of ethical approval in late 2012. It meant that 20% of participants were excluded. It is an acceptable loss as it meant that physical collection of GP records which could only be done by the researcher within the ethical approval time period was maximised.

- Nottingham based practices were contacted in descending order of number of consenting participants. Using this approach did not exclude participants.
- Two very large practices covering 29 and 35 patients respectively declined to take part despite ethical approval, supporting references from faculty members and letter of access being provided. The practice managers' requested that consent should be within in the last 3 or 6 months. It was explained that the study author was currently not permitted to contact any participants again.
- This meant a loss of two practices and 7% of participants (n=64) who had consented for their EMRs to be used but their respective practices declined. This study needed both patient consent and practice cooperation for EMRs to be collected. There was no information provided by these two practices on what software systems these practices used.
- All 28 of the remaining Nottingham practices were contacted and 717 electronic GP records were collected. They were securely downloaded and transported using the latest encryption technology and are stored in a highly secure encrypted data area at the faculty.

For ethical approval for the Gedling Cohort Study please see appendix 8.4
For the letter of access to primary care practices please see appendix 8.5
For a sample letter to General Practitioners please see appendix 8.6

2.7 The GP Systems Found at Nottingham Practices

Initially, arrangements were made to visit smaller practices to capture the EMR export process in a manner as if a patient was moving to another practice. That was the requirement and scenario presented to practice staff whenever a visit was made. The rationale for visiting smaller practices first ensured a confidence building process could occur by checking the steps used from practices (by GP software system) visited and compared to any subsequent ones. This process allowed a consistent and verifiable process to be documented irrespective whether the practice was large or small as the point was being able to extract a complete and reliable EMR version. This meant that the data extraction steps for each GP software system encountered were the same across practices. There was no risk of bias from smaller practice vs larger practices as keeping the same steps for a complete EMR meant good standards of practice early on to eliminate or minimise effects from differences in practice size and local GP software settings. We followed up participants at their most recently reported practice. This was the same as the practice they reported during their most recent follow-up. No participant contacted the faculty indicating they had moved practice. If a participant had moved practice it was normally possible to still extract their EMR until the point they had left. This was also the case for deceased participant EMRs which were still extracted. The objective during EMR data collection was to ensure all EMRs as possible were collected. There was no specific number for high reliability and validity powers as the objective was to create a new database similar to THIN from all the collected EMRs irrespective of the number collected. An output format, essentially a type of organised file, readily importable into a range of statistical software packages was required. The output format needed to be as universal as possible to manipulate and process electronically. The potential approaches to processing the EMRs effectively could then be explored once a range of raw (extracted from practice

only) GP records were obtained. Five software packages were found to be in use at the practices contacted; systmOne, iSOFT, Emis LV, Emis Web and Emis PCS.

This is shown in table 4 which gives an overview of the systems encountered during the data collection process by software vendor:

Table 3 GP Software System by Practice Visited During 2011-2012

GP Software System	Number of Practices (N	Total Number of Patients (N)	Percentage of Patients (%)
systmOne	17	489	68
iSOFT	2	144	20
Emis LV	7	68	9
Emis PCS	1	11	2
Emis Web	1	5	1
Total	28	717	100

Table 4 shows that systmOne was the most commonly found GP software system, followed by iSOFT, Emis LV, Emis PCS and Emis Web respectively.

2.8 Rationale for Choice of File Export Format of EMRs

Whilst all EMR systems store information using a highly-structured database, it is not usually possible to extract data from the database without installing additional software on GP PCs, or applying to the software provider to create an extract using the online backup system (as is done for the THIN data collection). Both are impractical for small-scale EMR data-collection exercises such as this Gedling cohort study. It is therefore necessary to work with the export facilities exposed in the software user interface at GP practices. These are designed to produce human-readable output, which means that individual items of data are no longer isolated in their own variables and or database columns. These outputs must therefore be processed into a format suitable for analysis.

2.8.1 The Export Options Available in the EMR Systems Encountered at GP Practices

Prior to any visits GP practices were asked about their EMR software system and where possible any online documentation available was investigated to get a familiarity of the GP software system in question beforehand. The steps for data extraction were explained by practice staff and documented and ensured that consistent steps were used across practices thenceforth. To facilitate data

processing using software, the following output EMR formats were considered suitable for this project (in order of preference):

- Structured text data files (e.g. fixed format, comma separated values)
- Human readable plain text
- PDF (Portable Document Format) is proprietary but electronically readable using freely available programming libraries

The reason for these preferences is due to text files and rich text files being designed primarily to be human-readable in their format and are straightforward to parse ("read") in software as software libraries have long standing existing built-in support to read such files. PDFs and bitmap images are also designed to be human-readable when displayed, but are encoded in a format that makes them difficult to parse in software without the use of software support libraries. Table 4 shows the export options available in each GP software package.

Table 4 - Export Options for Electronic Patient Records (EMRs) at GP Practices Encountered During Visits to Collect Patient Records

GP Software	Export File Format Options	Selected Format
systmOne	• Rich text file • Text file • comma separated value file (GP Practice level customised)	Text file
iSOFT	• PDF	PDF
Emis LV	• Text file • Jpeg/Bitmap images	Text file
Emis PCS	• Text file • Comma separated value file	Text file
Emis Web	• Rich text file • Text file	Text file

A plain text file was the most common file available in which to export EMRs across all GP software systems and is supported by near universal support in software parsers that read in data files. Jpeg and bitmap images were not used as they are difficult to read into software parsers to extract textual data and their translation can introduce unwanted and erroneous characters. The PDF file format was used as an export format from the iSOFT GP software system as that was the only format that the practice indicated they could provide for a complete EMR where a patient moves practice. These PDF files will need a suitable parser to translate them into text files. This will be outlined in chapter 4.

2.9 The Potential Approaches to Processing Exported Electronic Medical Records and Extracting Data

Once an exported plain text or PDF copy of the electronic record has been obtained from the GP practice, the data from the EMR needs to be extracted. Data extraction refers to the operation of retrieving data from a source system/file for further use (109). Often the source is called the 'raw' file as it is unprocessed as no data is yet extracted from this file. Obtaining a raw file is the first key step towards utilising the potential data stored within an EMR for epidemiological analysis (110). There are many approaches available to retrieve electronically stored EMR data (111–114). Several potential approaches to extracting the required data were considered. These were considered based on how well they would work with the file format of the collected EMRs, cost of purchasing and/or using the chosen extraction process and the existing skill set of the researcher to gain a good enough understanding to an advanced level so as to define in detail the extraction requirements. In addition, the selected extraction process should have also allowed the output to be validated for completeness compared to the raw EMRs and how flexible the process would be for the number of EMRs obtained. Finally another key requirement was how well this process could be replicated and distributed to fellow researchers in future

research studies looking at follow-up of EMRs in their respective cohorts. There is no consensus as to what the best and most valid and reliable method or set of processes are in relation to the processing of raw EMRs (115–117). Moreover the discussion around this becomes further fragmented when the diverse nature of different EMR vendors and their design and layout are considered (118). Even the same GP software vendor can produce several editions of GP software each with its own unique layout and structure, examples being Emis LV, Emis Web and Emis PCS. At this moment in time it is indeed a national and global phenomenon to have EMRs in an increasingly electronic format (119,120). However questions around the best format for an extraction and the best methods with which to do that extraction remains unanswered and there is no current consensus (121–123). In this study the best approach was to consider looking at the current extraction approaches available bearing in mind the above requirements and also in terms of cost, repeatability, completeness of extraction, future distribution and re-use. Therefore the following approaches were explored:

Manual copying and pasting: This involves looking at the data visually using suitable viewer software and for each electronic patient record, using the copy and paste functions to transfer the relevant information into an alternative file. However, this is a far too time consuming process for the number of patient records and information could be missed due to human error.

Semi-automated searching: This option involves using the 'Find and Replace' functions in software packages used to view the extracted records in order to find specific words/sentences quickly and then copying that piece of data into an alternative file for further processing. This approach is quicker than the above copying and pasting to get to the specific areas of the record and there is a less likelihood of missing data. It combines the above approach of using copying and pasting with the computer's ability to find data segments quickly. As a result,

there is much less manual copying of data then than the first option but it's still too time consuming to be a viable option even for small-scale number of EMRs.

Tokenisation: This approach has a wide variety of applications ranging from encryption to Internet indexing when retrieving an electronic document. Using tokenisation a text file is split into sentences using so-called 'segmenter' software and then each sentence is further subdivided into whole words ("tokens") using a tokeniser. The frequencies and positions of each token within a file or group of files are then used to build a searchable index allowing them to be located rapidly. This approach works well in contexts such as Internet indexing (124), where the aim is primarily to locate (sections of) electronic documents of interest. However this does not enable the data to be used directly because single tokens are seldom atomic in terms of their meaning: a Body Mass Index (BMI) measurement, for example, naturally comprises the measurement itself and a unit or other text identifying it as such, therefore extracting BMI recordings would require the application of complex logic to determine which combinations of tokens represent a single measurement, and which relate to other numeric quantities.

Natural Language Processing: Natural language processing (NLP) is a field of computer science utilizing linguistic analysis to gain "understanding" from unstructured free-text e.g. ability of Internet search engines such as Google or Bing to interpret complex queries and return most relevant results (125). The overall goal of NLP is to translate natural human language into a structured format each with a set of choices for its targeted value(s) and then order the retrieval information into subcategories or even query for the presence or absence of a finding (126). For example a recent study outlined how they studied discharge summaries using a diabetes classifier that given a discharge summary utilized discourse evidence, both direct utterances of diabetes and diabetes medication utterances to identify reports associated with diabetes (127). However a major downside is that the NLP extraction approach needs

time to build its 'learning' over a period so that it can be 'trained' to detect syntax in an unstructured environment (127). Missed signs and the resultant misclassification occurs because of spelling errors, hypothetical clauses, and syntax causing false positives in retrieving quality data (125).

Regular Expressions: As noted above, most EMR export formats such as plain text files are designed for human readability. The underlying data however is held in a highly structured database, with human-readable versions of the data generated automatically when required. The pattern and positions of individual pieces of data within the exported files are therefore highly predictable. Regular expressions are an extremely flexible way to define patterns which can be used to identify individual pieces of data within a text file, even based on their positions within the file and relative to the formatting characters or labels surrounding them (128). However a limitation of regular expressions is that it is difficult to automatically generate regular expressions that can accommodate unseen patterns so they cannot be used in any incremental learning in the same way as natural language processing (129). Regular expressions are selected as the extraction approach due to several reasons. Further rationale behind this choice is given next and is primarily because it is free to use approach, the researcher already has an existing skill set to develop the expressions, it can be used on several different software platforms without extra cost or copyright and is highly sensitive to specific patterns. This lends itself very well to the structured layout of an EMR.

2.10 The Selected Data Extraction Mechanism: Regular Expressions

Using regular expressions for the task of extracting usable data from the EMRs was considered the best approach after reviewing the available options. Analysing EMRs with regular expressions will allow the development of a series of consistent patterns and thus by comparing an EMR to this set of patterns the

computer can determine what is necessary to be cut out (extracted) and where to cut out from and up until; highly specifically and accurately. Regular expressions are designed to work especially well with text files and this was also the most popular export choice available at GP practices as well. There is a natural compatibility between the exported plain text files from GP practices and regular expressions. There was one exception where the iSOFT GP system only offered PDF exported EMRs. This exception was also amenable to searching using regular expressions once a plain text file is generated from the PDF EMR meaning that regular expressions could handle a range of EMR file types. Further information on how this is accomplished is provided in chapter 4.

2.11 Strengths and Limitations of Extraction Method

A key strength of regular expressions is its ability to identify individual pieces of data based on patterns and/or positions and sequence in the files relative to the (predictable) automatically-generated text surrounding them. This feature is crucial as each GP software vendor has a specific layout for their respective EMRs. A limitation of regular expressions is their lack of flexibility in that they even a small variation that breaks the specific pattern a regular expression is searching for will result in no data being extracted.

2.12 Previous Studies that have Used Regular Expressions to Extract from Electronic Medical Records

The concern around extraction methodology in existing studies is around how an extraction approach can be sure that the data extracted is actually providing the full and accurate view of what EMRs contain. For example one study concluded from a systematic review that there is little consistency or potential generalisability in the methods used to assess EMR data quality (130). There are existing studies that have used regular expressions in EMR data extraction and have demonstrated validity and accuracy using their respective gold

standards. The use of regular expressions can compare favourably for data extraction from an accuracy and completeness point of view (131). This study will draw upon these to formulate the extraction methodology and provide a rationale for the study design next.

2.12.1 Formulating an Extraction Methodology Using Regular Expressions and Manual Review and Setting a Gold Standard

Several previous studies that have extracted data from EMRs have utilised manual review as part of their gold standard to evaluate effectiveness of their extraction methods. For example the earliest study found to use regular expressions and EMRs was in 1996 and used the following steps that formed their methods (132):

1. Establish a model or definition of the concepts to be extracted e.g. drug, dosage etc.
2. Prepare the data for training and testing
3. Create an Info Extract module (software or algorithm)
4. Processing the data to extract dosage references
5. Scoring results

The same set of steps formed the basis of a separate study 12 years later in 2008. The authors of the 2008 study stated that their precision and recall were comparable to the results reported in the 1996 study and cited this paper due to the similarity of the needs in extracting EMR data to a high precision of accuracy and that the 1996 study formed the basis of their methods (133). They also found their manual evaluation showed a precision of 94% in the extraction of medication information and was based on a gold standard created by cooperation between two reviewers of extraction results. Several further studies applying regular expressions and manual review for evaluation of randomly selected EMR extraction found similarly high precision results post extraction (134–136). Even in studies that did not use regular expressions; have still used manual review in combination with their automated extraction mechanism. For

example a study that used an algorithm developed in Microsoft Excel's programming module combined with manual review concluded that their extraction method was 99% accurate on a random sample of 500 dosage texts from the CPRD (137). The findings from previous studies show that involving manual review with automated extractions from EMRs is a useful measure to establish a gold standard of checking and verifying accuracy and completeness in extracted data files.

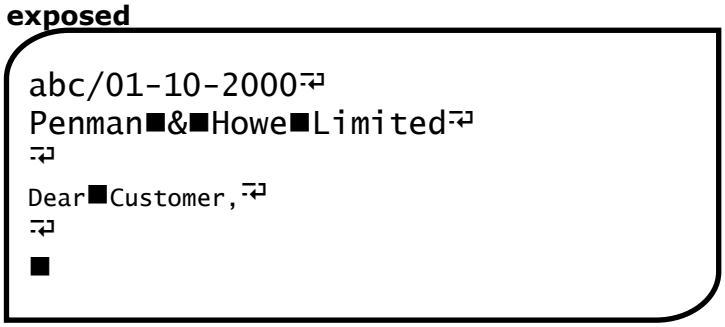
2.12.2 Accuracy and Completeness and Establishing the Gold Standard for this Study's Extraction Methodology

This study will verify accuracy and completeness of the extraction software module in an iterative and staged process where during software development and testing of regular expressions there will be an examination of line by line data produced for each software vendor using three randomly selected EMRs per vendor. During software testing if there is data missing then the regular expressions will be continually fine-tuned until working properly so that all the EMR data is extracted into the data files. The objective is to ensure all the data is extracted electronically and stored in accessible data tables for statistical analyses. It is possible to do this as it is a relatively small scale study with a specific cohort whom we know things about in advance from previous questionnaires and where data sought was structured. It is also limited to practices from a specific geographical area of Nottinghamshire which may have an effect if many practices use the same GP Systems of Choice (GPSoC) scheme to source their IT hardware meaning the same computer hardware and configuration is common because of the same supply and service chain. For further information please see chapters 6 and 7 where the EMR extracted data results are presented and discussed.

2.13 Explanation of Basic Regular Expression Notation

Regular expressions are the programming tools used to specify patterns of what to extract from EMR plain text files. Regular expressions are built from their own unique programming language and this is described and explained next. The following regular expression notation is used throughout this study. For example, suppose a text file contains the customer reference number for a fictitious company at the very top of a letter as shown in Figure 2 :

Figure 2 a set of standard text with corresponding hidden characters exposed



Characters Regex Can Understand:

↵ carriage return & new line

■ space

⇊ tab

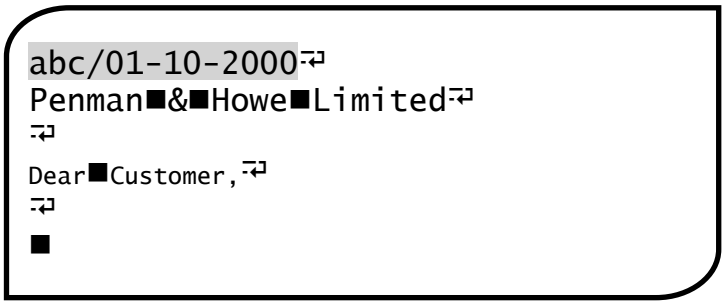
The fictitious reference number is contained in the first line and followed by other text. However it still can be uniquely differentiated from the rest of the letter and anything in the letter that might be similar to the reference number as well like a date for example. To enable this to happen (in an electronic version of this letter) the computer must be instructed using programming exactly what the reference number looks like so it can be identified and extracted from wherever it is found. This programming notation to specify what the target text looks like is called regular expressions. Therefore regular expressions (regexes for short) need to be highly specific so that the regex extracts only what is needed in a computerised text field. In the example above, to extract the customer reference number the following steps would need to be understood by the computer and supplied using regex notation:

- Look for the following characters in this specific structure:

- The Letter ‘a’ followed by ‘bc’ followed by forward slash
- Then followed by digits 01 followed by single hyphen
- Then followed by digits 10 followed by single hyphen
- Then followed by digits 2000
- Finally followed by a carriage return and line feed

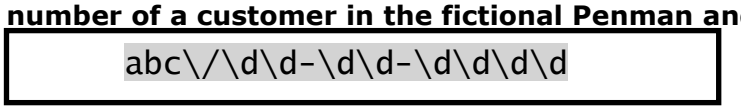
The file is sequentially searched, and once all the above criteria are met then what is found is the letter’s reference number, for example in Figure 3 :

Figure 3 target area for regular expression shown in grey highlight contrasted with rest of the standard text



This would be expressed as the following regular expression to locate the needed customer reference number in Figure 4 :

Figure 4 A single regular expression targeted to locate the reference number of a customer in the fictional Penman and Howe limited



The contents can then be copied to the computer system’s memory and output later on into a more structured output file. Such specific steps require a special notation to explain to the computer system exactly how to do such steps. By examining the hidden electronic characters as shown figure 4 and 5 appropriate regexes can be written to identify an array of possible targets ranging from large sections of an EMR to specific words and sentences.

Table 5 is basic list of regular expression meta-character notation in Microsoft's C# Programming environment (138):

Table 5 list of regular expression meta-character notation in Microsoft's C# Programming Environment

Meta-character	Behaviour	Example
*	Matches the preceding character or subexpression zero or more times.	to* matches "t" and "too".
+	Matches the preceding character or subexpression one or more times.	fo+ matches "fo" and "foo", but not "f".
\b	Matches a word boundary; that is, the position between a word and a space.	er\b matches the "er" in "never" but not the "er" in "verb".
\d	Matches a digit character. Equivalent to [0-9].	In the searched string "12 345", \d{2} matches "12" and "34". \d matches "1", 2", "3", "4", and "5".
\r	Carriage-return character	Hidden character
\s	Any white-space character. This includes space, tab, and form feed	Hidden character
\t	Tab character	Hidden character
\n	Newline character	Hidden character

Table 5 is not an exhaustive but provides the basics that will be used to produce the expressions. A regular expression is evaluated similar to an arithmetic expression from left to right.

2.14 Programming Languages Available for Writing Regular Expressions

Regular expressions generally have the same basic notation across the various widely used advanced software languages including Microsoft C#, Java, Delphi, Python and Perl. It is helpful to use such comparatively advanced programming languages that supports and implement regular expression standards as they are feature rich in developing front end software modules so an end-user can use point and click to use any software produced. It is not however absolutely

essential that such programming languages are utilised for writing regexes. They can for example be written from Disk Operating System (DOS) prompt that comes preloaded with Microsoft Windows. They are however a necessary platform to harness the full potential of regular expressions, such as extracting data to a specifically created data object or writing comments along with the regular expressions so another person can read and understand the regular expression code. Software languages also have the ability to be used from within familiar Graphical User Interface (GUI) tool that we associate with products such as Apple Corporation's Mac OS X and Microsoft's Windows. This makes the process of writing complex regexes and examining their results less time consuming and much easier to understand. There is no right or wrong programming language to use for regular expressions as the notation is very similar across them. Considering the time resources available, the researcher's own programming experience and quality of regular expression implementation, the decision was to use Microsoft's C# programming language within which to write the Extraction Software (ES). There are lots of regex libraries in different software tools; however C# is best for widest accessibility and compatibility. The ES will in effect act as wrapper which will allow end users to work with a GUI computer program to process the EMRs by pointing and clicking on a screen.

2.15 The Microsoft C# Programming Environment

Another reason for choosing Microsoft's C# in this project is because it produces portable code which can be run on a variety of platforms and includes a package that contains a range of tools specifically for regular expression tasks. Thus it is possible to invoke the regular expressions package and usefully apply it without creating a regular expression interpreter from scratch. There are also a large number of books and online resources explaining C#'s regular expression toolset. So it means on-going long term support for the language and usage.

2.16 Summary

The Gedling cohort is being used for collecting EMRs from consented patients and THIN is used as a template to create the new health database based on the extracted data. In total five different GP software systems were found and a total 717 patient records were captured. There are many different file export options available in each GP software system. Plain text files were best for being a commonly used file in all GP software systems and existing wide support for its interoperability with software development tools to read in and process the EMR plain text files. The best mechanism to extract the data from the EMRs was determined to be regular expressions because of their strength in being able to extract highly specific patterns. Previous studies that have used regular expressions with EMR data extraction have involved some form of manual review. This study will verify accuracy and completeness of the extraction software module in an iterative and staged process where during software development and testing of regular expressions there will be an examination of line by line data produced for each software vendor using three randomly selected EMRs per vendor. During software testing if there is data missing then the regular expressions will be continually fine-tuned until working properly so that all the EMR data is extracted into the data files. The objective is to ensure all the data is extracted electronically and stored in accessible data tables for statistical analyses. The resulting database files will then be subject to independent blind review by two checkers who randomly select EMRs to check that data in raw EMRs has all been extracted. It is possible to do this as it is a relatively small scale study with a specific cohort whom we know things about in advance from previous questionnaires and where data sought was structured. Once it passes the iterative testing and verification stage our Gold standard of all EMR data extracted into the database as being fulfilled. For further information on methods and strengths and limitations please see chapters 6 and 7 where the

EMR extracted results are presented and discussed. Chapter 4 focuses on extracting the data using the chosen approach and explains the design of the data files that together form the database. Chapter 5 provides a detailed description and explanation of the extraction approach for each GP software system encountered during EMR collection.

3 Overview of the Conversion Process of Gedling Cohort Data Sets into a New Database

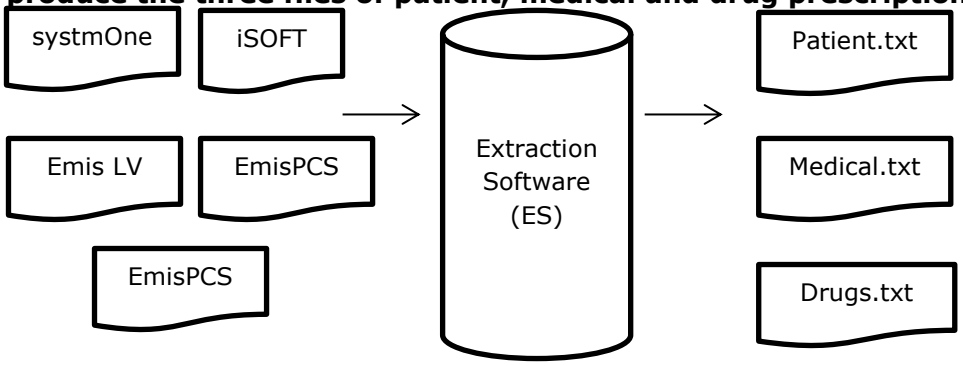
Aims of this chapter:

- Describe the various stages that were needed to extract data from the EMRs
- Describe the extracted data files
- Explain how dosages found in the extracted data files can be extended
- Explain reasons why the extracted data files are a good design for research

3.1 Introduction to the Extraction Software

Once the collected EMRs were collated together in an electronic folder it was necessary to find a solution to extracting the necessary data from these raw unprocessed files text files. This chapter describes how this was achieved by using regular expressions to 'cut out' any data that matched the criteria specified. The regular expressions were incorporated into a specially developed Extraction Software (ES) which read in each EMR and applied the relevant regexes depending on the EMR originating GP software system encountered finally producing the extracted files. This process is shown in Figure 5 :

Figure 5 The different EMRs interacting with the extraction software to produce the three files of patient, medical and drug prescriptions

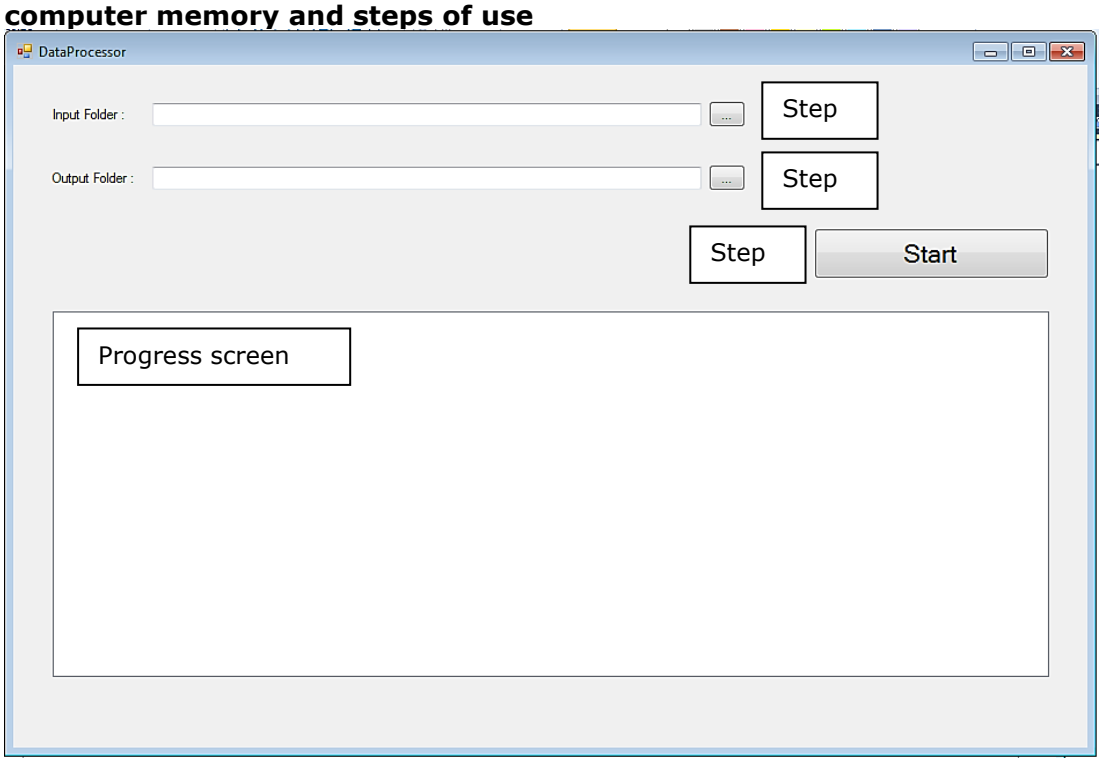


3.2 Steps Users Take to Utilise Extraction Software

The user process needed to utilise the extraction software is shown below in Figure 6 . Need to be more clear all the text files just in the same folder? Or how is the data stored. The extraction software is used by following the steps below:

- Step 1: select source folder where raw EMRs are stored: "Input folder"
- Step 2: select destination folder where output files will be stored: "Output folder"
- Step 3: select the button: "Start"

Figure 6 User interface of Extraction Software when running in computer memory and steps of use



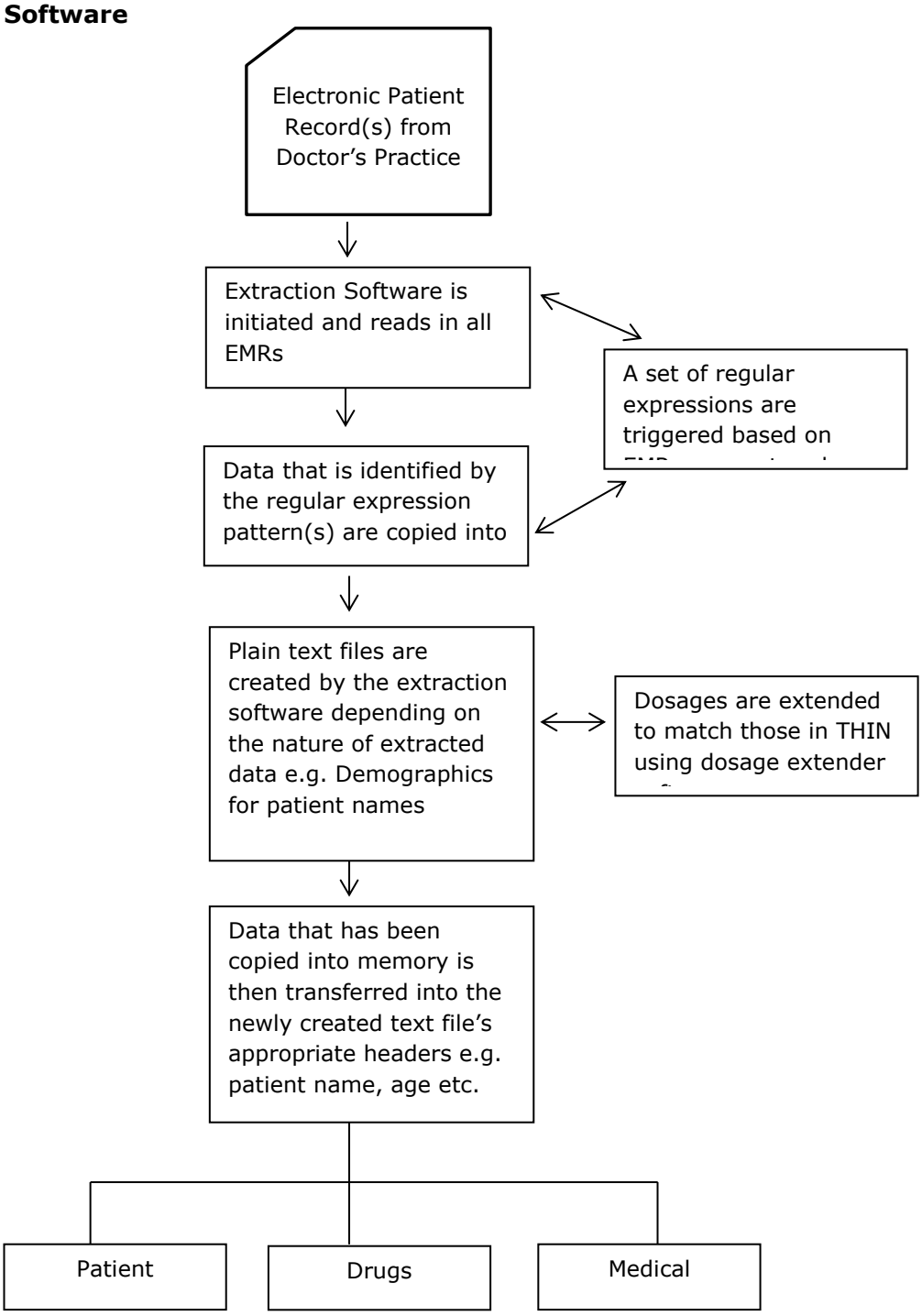
Post processing three plain text files are created in the destination folder and a message to this effect is displayed in the progress screen area. If the destination folder is then opened; the end user would see these three extracted files ready to be used in statistical software e.g. Stata. The three files are based on the THIN tables of 'patient' which covers demographics, 'medical' which covers medical events like symptoms and diagnoses and 'therapy' covering

prescriptions data. Another key strength of the design is that it is as simple as possible using interface features that most users would be familiar with such as a button and two drop down lists. This means that users do not need special training to utilise the software. Another advantage is that the software is that it can be ubiquitous as it is designed to portable and run on any PC and Mac without further software modifications to a computer system. The source code is attached on CD in appendix 8.7

3.3 The Data Extract Stages

All EMRs are processed with the same overall approach though different regular expressions depending on the originating EMR system which the extraction software detects beforehand. This process is outlined in Figure 7 :

Figure 7 Algorithm for extraction of EMR data used by the Extraction Software



3.4 The role of primary and secondary regexes to do data extractions

There are two levels of regular expressions which are triggered at different times and applied separately depending on the originating GP software system downloaded from. A primary regex refers to regular expressions that detect the sections of an EMR being analysed. This indicates to the extraction software the boundaries of an EMR thus allowing it to split an EMR into subsections. This is useful because it allows a greater specificity of patterns to occur in subsections where there can be greater uniqueness of patterns compared to a whole EMR.

3.4.1 Applying Primary and Secondary Regular Expressions

Initially sections were manually marked out of several paper versions of EMRs from all the different types of GP software systems. The paper records were also used to start sketching out regular expressions on a whiteboard. This formed initial designs of primary and secondary regular expressions. Using the software Notepad ++ it was possible to expose the hidden computer characters that hold the EMRs in their layout on the screen. Primary regular expressions were needed so as to split a record into its constituent parts according to the structure. The aim of this initial split was to break up data types according to their overall position in the document. This avoided problems like preventing regexes targeted at (e.g.) prescription dates from picking up the patient's registration date in error. The initial extraction split an EMR into these three EMR subsections covering the areas of patient demographics, prescribed drugs and medical encounters. Next regexes were coded to mark out the following main areas to break out of an EMR to begin the extract process. These newly divided subsections can be thought of as mini pieces of the whole record and these become the basis for the second stage 'secondary' regexes. For example initial split and secondary extractions:

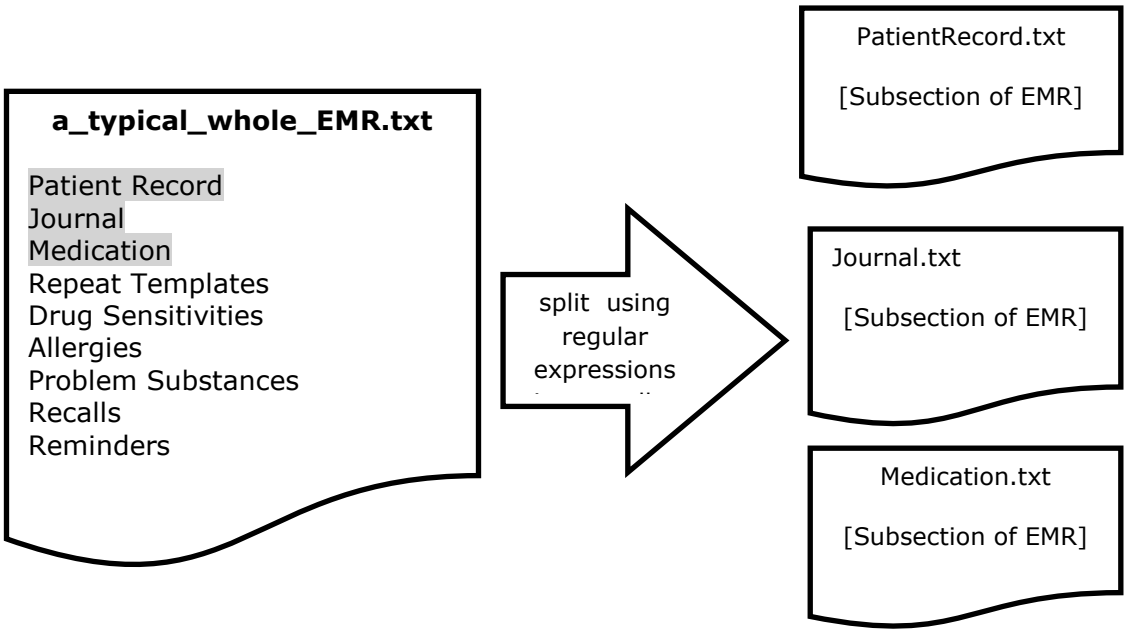


Figure 8 - Regular expression highlights the higher-level sections in a whole record that then form the three key subsections

Once the file is split as shown in Figure 8 then secondary regexes were applied to each subsection for further extraction of data rows and columns that form text files of patient, medical and drug prescriptions. An example covering demographics is seen in Figure 9 extracting from individuals' PatientRecord.txt into the demographic information of all patients in a single file 'Patient.txt'.

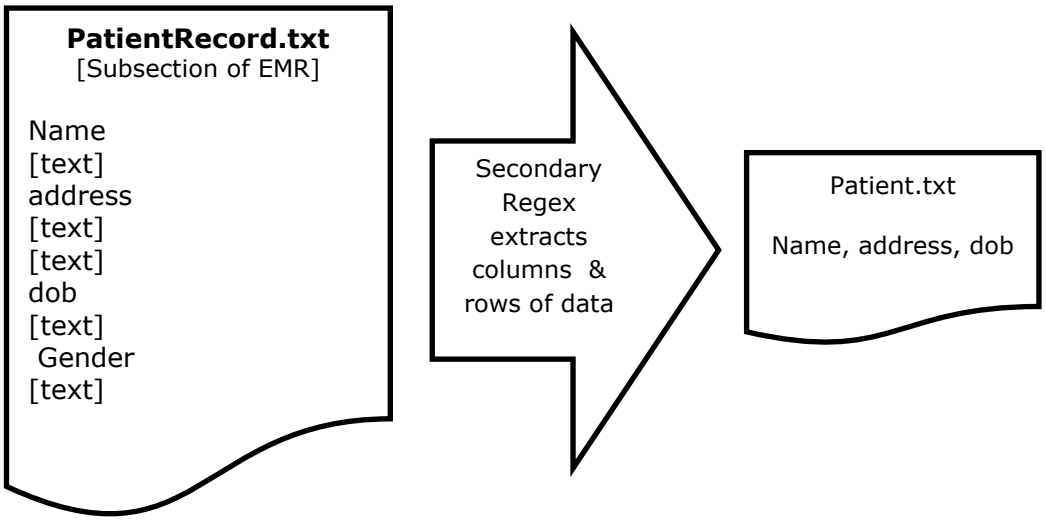


Figure 9 - secondary regex extraction from a subsection containing demographic values

Although developing primary and secondary regexes were created on paper to begin the process they had to be done electronically before it was possible to finalise the regular expressions used in the deployed version of the extraction software. Chapter 3 previously introduced methods to ensure accurate extraction. Prior to manual review of the extracted data the researcher also had to ensure data was being extracted properly before the files are created. It was achieved by iterative testing of regular expressions both primary and secondary expressions by viewing the randomly selected EMRs on screen coupled with an on-screen regex tester called 'Regex Buddy' (from Just Great Software Co. Ltd). This allowed the researcher to create a test environment so that iterative testing of regular expressions could be applied to the EMR data until it was fine tuned to ensure data was extracted properly. The files post processing were then manually reviewed by two independent reviewers to ensure data was extracted accurately. Further information on EMR extraction methods and results are covered in chapters 6 and 7.

3.5 Rationale for the extracted data files structure

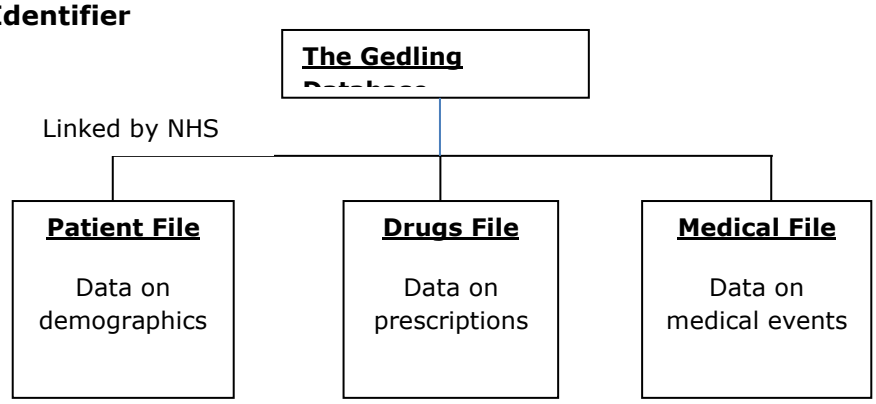
As stated the output format of files was designed to be similar to the record format used in The Health Improvement Network database (THIN). THIN has a widely-used and well understood data format which accommodates the overwhelming majority of data types stored in EMR systems in general and is therefore suitable as a template for the final data output format. The THIN database has a very specific structure as it has been designed to hold data in a clean and organised way. By being properly designed it means that the analysis of data held in this format will be accurate, complete and allowing statistical testing and produce relevant data sets. The final output files were created in a plain text electronic format and contained the following information:

- "Patient.txt" - Patient information e.g. Name, Address etc.

- "Medical.txt" - Journal details (contains all Read codes- medical diagnoses)
- "Drugs.txt" - Prescriptions given to patients, name of drug, dosage, instruction

A full description of each output file is given in the next section. Figure 10 gives an overview of how they are related to each other:

Figure 10 - Structure of the Gedling Database Data Files linked by NHS Identifier



All three data files are linked by each participant's unique NHS identifier. Each file is explained in further detail in the next section.

3.6 Extracting information about demographics

This refers to area and personal related information about patients. All relevant demographic in each of the systems was extracted. It was also possible for certain pieces of demographic information contained in one GP software system and not another. For example 'RPP/Rural' referring to miles travelled if patient was classed as living in a rural area, was present in Emis systems but never in systmOne. The basic demographic data that was common to all GP software systems and duly extracted was:

- NHS identifier
- full name
- age
- address
- date of birth

Table 7 shows the full listing of fields contained in the demographics file.

Table 6 - Full listing of patient table fields, character type and description of patient fields

field	character type	description
name	alphanumeric	full name of patient
nhsid	alphanumeric	unique identifier used by the NHS
dob	DD/MM/YYYY	date of birth
address	alphanumeric	address of patient
registeredgp	alphanumeric	registered GP of patient
gender	M | F	sex of patient as M:male or F:female
registrationdate	DD/MM/YYYY	date patient registered at practice
usualgp	alphanumeric	a patients usual GP if indicated
postcode	alphanumeric	patient's postcode
ethnicorigin	alphanumeric	the patient's ethnic origin
maritalstatus	alphanumeric	the patient's marital status
dispensing	alphanumeric	whether practice is a dispensing one
tradingpartner	alphanumeric	NHS trading partner information
recordsat	alphanumeric	records at – location of records
hospitalno	alphanumeric	hospital number
residentialinstr	alphanumeric	residential instructions palliative care
walkingquarters	alphanumeric	walking quarters data
regstatus	alphanumeric	registration status
language	alphanumeric	language(s) spoken by patient
patienttype	alphanumeric	patient type permanent or temporary
advocacyneeds	alphanumeric	advocacy needs
transportneeds	alphanumeric	transport needs
carer	alphanumeric	carer details
emisno	alphanumeric	emis number
rpproadmiles	alphanumeric	rural road miles travelling by patient
specialdist	alphanumeric	special district information
placeofbirth	alphanumeric	place of birth
englishspeaker	alphanumeric	english speaker
deceased	alphanumeric	deceased
dateofremoval	DD/MM/YYYY	date of removal
pdsregisterdpractice	alphanumeric	personal demographic service practice
registrationtype	alphanumeric	registration type of patient
medicalrecords	alphanumeric	medical records
Pharmacy	alphanumeric	pharmacy details attached to patient
appointments	alphanumeric	appointments
attendance	alphanumeric	attendance rate out of 100%
visits	alphanumeric	visit rate out of 100%
dnas	alphanumeric	did not attend rate out of 100%
practiceid	alphanumeric	unique practice identifier
collectiondate	DD/MM/YYYY	collection date of electronic record

Table 7 shows all the demographics fields that were extracted from the EMRs covering all GP software systems encountered these were extracted into one single unified demographics file. Chapter 5 explains and describes how this was accomplished for the EMRs covering all GP software systems encountered.

3.7 Extracting information on medical events

This refers to the recording of medical care in terms of diagnoses that the patient has received. In most GP software systems these encounters are encoded by Read codes or their descriptions. Table 8 shows the full listing of fields contained in the medical file and it means that for every patient the following medical event data is extracted:

- Read code Text – this refers to the name given to the Read code itself. It is a description in English of what the Read code represents.
- Read code – This is the Read code itself and represents a 5 digit unique identifier of the therapy prescribed. Some GP software systems contain the Read code text and not the Read code. They can also use their self-created codes loosely similar called local codes.
- Read code Free Text – This refers to the supporting text that follows the Read code itself and is ad-hoc text inserted by the doctor or practice themselves. It is optional and may not be present each time.
- Date of recorded Read code – Each medical encounter normally has a date associated so it becomes possible to define a pattern that matches the date the patient received the therapy indicated by the Read code.

Table 7 - Full listing of medical table fields, character type and description of medical-related fields

field	character type	description
nhsid	alphanumeric	unique identifier used by the NHS
date	DD/MM/YYYY	date of Read code recording in EMR
Readcode_text	alphanumeric	text associated with Read code
Read code	alphanumeric	Read code recorded onto EMR
free text	alphanumeric	other text added manually/GP software
staff name	alphanumeric	any staff names indicated EMR record
vendor	alphanumeric	originating GP software at practice
readcode_text_local	alphanumeric	text describing Read code from EMR
measurements	alphanumeric	any measurements recorded with event
row number	alphanumeric	row number

Table 8 shows all the medical events field that were extracted from the EMRs covering all GP software systems encountered these were extracted into one single unified medical file. Chapter 5 explains and describes how this was accomplished for the EMRs covering all GP software systems encountered.

3.7.1 Extracting free text in the electronic medical records

It is possible that text is encountered that is not present in the same format and structure as in all the EMRs of a specific GP software system. Generally some external function inserts the text such as the computer system itself or the doctor or practice staff. It is still important to retrieve such data so that it can be humanly assessed as to how useful and what insight can be gained. This type of text is called free text and is an increasingly common phenomenon associated with the increasing use of electronic medical records (139). In this study if the computer parser encountered text that did not match the regular expressions designed then the text read into memory was placed in the free text column for further analysis.

3.8 Extracting information about drug prescriptions

This refers to the recording of prescriptions of drugs for each patient. In most GP software systems these encounters are encoded by Read codes or their descriptions of Read codes to indicate a coded prescription. Table 8 shows the full listing of fields contained in the drugs file and it means that for every patient the following drug data is extracted:

- Drug name – name of the drug prescribed by the GP
- Date – date the prescribed drug was recorded onto EMR
- Dose – Dosage indicated by the GP for drug prescribed
- Form – Form of the drug prescribed. For example tablets or liquid forms.
- Instruction – The number of times during a certain period the patient should consume the drug prescribed
- Drug Prescription – This refers to the complete prescription data when drug name, dose, form and instruction combined in the same data field.
- Vendor – This refers to the originating GP software that EMR was collected from at the practice level

A number of different approaches were tried to deal with prescription data to reach the end goal of type of drug, form, and dosage in separate fields. One method was to attempt to use positional matching as the drug name was to the right of the data extracted in line two after a date is encountered; however that did not work as it would result in some longer drug names being cropped. Another approach was to use the position of digits to represent dosage and then everything to the right would be the name and everything to the left would be the form as this pattern was present in many drug names. However this was not robust enough because some drug names also contain a digit as part of their name; for example 'E45 cream'. The approach then adopted to ensure maximum accuracy was to use second stage filtering. Each GP software system's EMR layout determined the variation to this second stage filtering

approach which is based on first removing the form and then removing the dose with drug name remaining. For example for the majority of EMRs extracted from systmOne meant that the extracted line 2 after the date line would be processed further to separate out the interesting parts needed into a cleaner format. This was done by taking the drug extracted from line 2 and then removing the form first by matching against a list of forms; then removing the dosage by matching through a list of possible dosage values and finally what remained after this filtering would be the drug name being sought. It allowed an isolation of data to produce name, dose, and form in the final extract. This method of filtering the drug related information was applied when processing all the GP software systems in the interests of keeping any name dose form consistent across all systems found. This process is explained in chapter 5 for each GP software system encountered. The lists used for form and dosages are in appendix 8.8. The list of forms was created using THIN's form definitions as a starting point and this process was checked externally by one supervisor. The forms list was then refined iteratively through software testing and data checking of the field values of forms and dosages. Where a form was not being extracted it was then added to their list of forms so it would extract correctly. The list of dosages were created from scratch by examining the output and adding and then added to as testing proceeded in the same manner as form definitions list. These definitions of forms and lists were used as a regex themselves by using programming to specify them to be used in that manner. This allowed the programming to use regular expressions' power of being highly specific and sensitive to patterns effectively for both forms and dosage patterns and also allowed an abstraction of patterns so that the forms and dosage regexes could be flexible to detect a range of patterns for both. Table 9 shows the full listing of fields contained in the prescriptions file:

Table 8 - Full listing of drugs table fields, character type and description of drugs-related fields and those fields created by dosage extender software to match THIN dosage fields

field	character type	description
nhsid	alphanumeric	unique identifier used by the NHS
date	DD/MM/YYYY	date of prescription recorded in EMR
drug_name	alphanumeric	name of drug prescribed
dosage	alphanumeric	dosage of prescribed drug
form	alphanumeric	form of prescribed drug
instruction	alphanumeric	prescribed instruction of drug
drug	alphanumeric	drug name, dose and form and instruction
vendor	alphanumeric	originating GP software at practice
dosval2	alphanumeric	see dosage extender study reference
mindos	alphanumeric	see dosage extender study reference
maxdos	alphanumeric	see dosage extender study reference
dosperiod	alphanumeric	see dosage extender study reference
conditional	alphanumeric	see dosage extender study reference
pair	alphanumeric	see dosage extender study reference
effdosval	alphanumeric	see dosage extender study reference
effmindos	alphanumeric	see dosage extender study reference
effmaxdos	alphanumeric	see dosage extender study reference
asdirected	alphanumeric	see dosage extender study reference
takemorning	alphanumeric	see dosage extender study reference
takeevenight	alphanumeric	see dosage extender study reference
withfoodmeals	alphanumeric	see dosage extender study reference
unitdosage	alphanumeric	see dosage extender study reference
dosageunit	alphanumeric	see dosage extender study reference

Table 9 shows all the prescription-related fields that were extracted from the EMRs covering all GP software systems encountered these were extracted into one single unified drugs file. Chapter 5 explains and describes how this was accomplished for the EMRs covering all GP software systems encountered.

3.9 Extension of dosages study

A further advantage of isolating the dosages from the drug name was that further extension by matching against a dictionary of THIN (The Health Improvement Network) dose definitions allows a greater ability to examine aspects of dosage not possible previously. This was achieved by integrating the extraction process with the lexicon cleansing process created by a colleague (140). Table 9 shows these fields as a shaded grey fields to indicate the new fields post extension processing.

3.10 Uniquely identifying a patient

The NHS identifier was used to identify patients from one another. To achieve this it required a modification of the program's regex patterns that initially processes and splits the record according to key headings and ensures that that the NHS Id for each patient was added or (stamped onto) to each subsection so that could then include the NHS Id with each piece of information that was further extracted and included inside the patient demographics file. For example, the data subsection for drug prescriptions also included the nhsid inside its own text file. This helps when doing drug prescriptions so they can then be related correctly to the specific nhsid of the patient in question.

3.11 The Gedling Database

The Gedling database is a collection of three files that have been described and explained above. It is small scale version of THIN with a specific cohort of participants whose EMRs have been collected from a wide range of GP software systems. The extraction software can be distributed and shared with other researchers and re-used without purchasing expensive software. A strength of the Gedling database is that created local health databases can help researchers gain a new insight into specific cohorts about which we may know some information beforehand but now it is possible to collect their EMRs and convert them into a database and use with existing statistical scripts and interrogate the health histories in a way that was hitherto possible with THIN, QResearch and CPRD health databases. A limitation of this approach is that it is not always possible to make generalizable conclusions for large populations using small scale health databases because the statistical power is not there. The Gedling database is not meant to compete with these large health databases but it is inspired by the notion that using EMR data it is possible to gain an understanding of health outcomes for a specific cohort. The next chapter describes in detail how the Gedling database is assembled together and provides

explanations and rationale for design decisions. It outlines the blueprints of the Gedling database by going through all the regular expressions used for data extraction.

3.12 Summary

EMRs from the various GP systems collected around Nottinghamshire were collected together and regular expressions were used to extract data. This was then used to generate three files covering demographics, medical events and drug prescription data. These three files make up the Gedling Database which used THIN as a template. All the data contained in EMRs is extracted into the three Gedling database tables. All EMRs are processed with the same overall approach though different regular expressions depending on the originating GP software system as each software system produce a unique EMR layout. The Gedling database is a small scale version of THIN with a specific cohort of participants' EMRs collected from a wide range of GP software systems.

4 The Extraction Software Module for Each Electronic Medical Record Type and Assimilation of the Gedling Database

4.1 Aims of this chapter

- Describe EMR structure by GP software system
- Describe the design of the regular expressions to extract data
- Explain the reasons for the design of each regular expression
- Explain the reasons for the design of these three extracted files

4.2 Background

The Extraction Software (ES) was designed to do all the necessary work of reading each EMR and deciding which series of regex patterns to apply depending on the GP software system. This means that the Extraction Software (ES) takes a collection of EMRs downloaded from GP surgeries and applies regexes to extract EMR data in an organised way to make epidemiological analyses and arguments. This chapter details what occurs in the mechanisms of the Extraction Software (ES) in detail view of all the regular expressions created and applied. During the extraction process each GP software system derives the following three files:

- Patient Demographics
- Medical encounters
- Drugs prescribed

These three files are produced for each type of GP software system. The Extraction Software then joins the files from each of these GP software systems together to create the same three files which are now in unified manner across all systems. For example the final 'Patient' file would contain the demographics from all the patients irrespective of which GP software system an EMR comes

from. The next sections describes the code for extraction software for each of GP software system including:

- Introduction to each GP software system
- Guide of how the EMRs are downloaded at the GP practice level
- Step by step process each regex takes by using fictitious EMR examples from initially splitting whole EMRs into subsections and finally into the derived files in terms of columns and rows in a text file.
- Description what the different regexes look like written in software and how this would be interpreted in real world human understanding
- Giving a rationale for the design of regexes created for each GP software system

These points will be outlined in the following section in order of GP software systems by descending order of participants:

- systmOne
- iSOFT
- Emis LV
- Emis PCS
- Emis Web

4.3 Producing the systmOne data file

4.3.1 Introduction to systmOne data extraction

SystmOne is a patient record system produced by a company called TPP plc. This section outlines how systmOne EMRs were extracted.

4.3.2 Exporting records from systmOne

The EMRs for systmOne were collected from practices in a consistent manner. Instructions on how to download systmOne EMRs are available in appendix 8.9.

4.3.3 Problems encountered when extracting data

Each software system provided its' own unique challenges in order to extract the data, for systmOne these include issues with data formatting in order to create an unified dataset across the different software systems and in addition issues with software programing and how they were overcome are discussed in more detail next.

4.3.4 Data format and layout issues

The first issue to enable standardisation of the date values across all files a decision was made to use the date format of "dd/mm/yyyy" across all date fields in extracted files, however within systmOne the date format of "dd mmm yyyy" (a combination of numbers, spaces and three alphabet letters to denote month). This was fixed by using C# code to convert all dates into a new format of "dd/mm/yyyy". Another issue found was the extracted data in either prescription or medical data into output files contained large amounts of unrelated text in the same columns due to the extraction algorithms not specifying what to extract properly. For example automated text on a practice's scanning system found their way into the medical file. This type of ambiguity issues were resolved by continual and incremental testing and checking of extracted values to original EMRs to ensure the post extraction files were as accurate as possible and did not contain incorrect information in the correct column of data. For example, in order to extract name, dose and form of

prescriptions from EMRs an initial approach was used where the position a drug name, dose and form occurred being separated by spaces was used as a pattern to break down a drug item into name dose and form. However this was not a good approach as it meant a heavy reliance on space characters and the output file examined had a lots of errors in it as no further intelligence being applied to break down what are key items to capture. This became apparent within the random testing because the space patterns were not consistent enough across all EMRs in systmOne. One method that was explored was to actually code names of drugs from customised list to extract items. However upon random testing and checking it was found that there was variation in drug names where the same drug was referred to its different brand names and in some places only chemical names recorded. For example salbutamol was commonly recorded as brand name ventolin and salmeterol would be called xinafoate or serevent in some prescriptions and therefore creating a list of drugs was not viable option. Finally a solution was found that would allow a consistent extraction across systmOne GP software systems. This was done by creating a series of regexes that would remove the form, such as (tablet, pill, liquid etc.). The form regex tree was formed after using the forms list found in THIN as a starting point and then added to by examining the extracted data for additional forms. These can be found in appendix 8.8. Following on from this process the dosages which were remaining were the drug name. The list of dosages were created from scratch by examining the output and modifying the list as iterative testing proceeded (see appendix 8.8 for dosage list) for the GP software systems encountered systmOne onwards. If dosage structure or forms were missed, iterative revisions to the dosage definitions were done.

4.3.5 Software programming issues

The first time the Extraction Software (ES) code ran in a PC's memory the code took it too long to execute and it was found that this was because the programming code was reading in a file and not releasing the file for further

processing and there this was correct in the computer code. Another issue found was that the C# code was not running properly on an Apple Corporation's Mac computer as an error message was being received that a file access could not be accessed. It was then sorted using an approach to combine the path structure so that Apple Mac's operating system and Windows could understand location of EMRs properly. It became apparent during software development that it would be very useful for analysis to have a collection date from a practice. Using the date from metadata that is part of a folder being created in computer during a GP practice visit created too many software errors during development and was open to be inadvertently changed or modified. The approach used was to ensure each collected folder also contained the date of extraction from the actual practice in the format of dd-mm-yyyy immediately following the practice name. New C# code was then added so that the collection date was formed from the date found in the folder name. It also contained new code so that a folder is not processed until the collection date is present in the folder name. Using digits and hyphens in the folder name was considered best as forward slashes confuses Windows PC's as they interpret this as part of path name variables so forward slashes are not allowed in folder names. This allowed a collection date for all EMRs to be calculated by the Extraction Software (ES).

4.3.6 Splitting the systmOne file into subsections

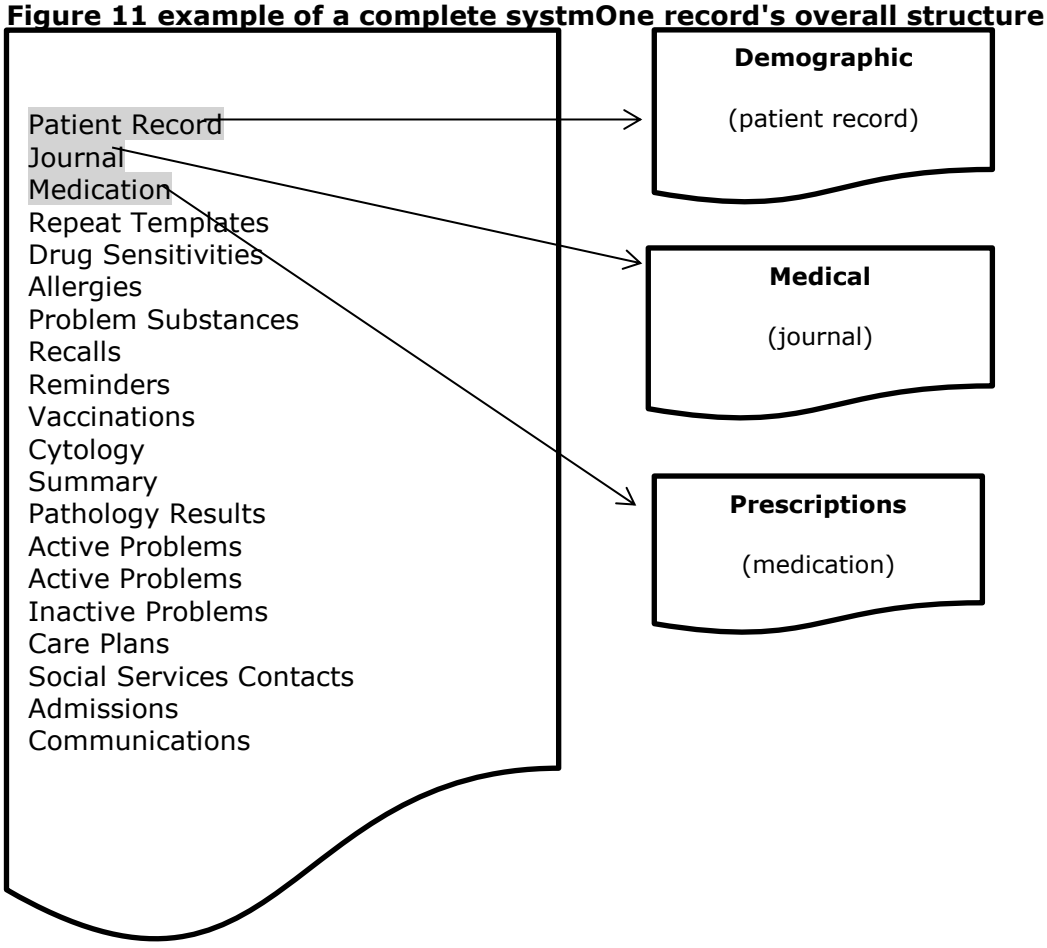
The extraction software applies the first stage of the regex to split the systmOne record into the following subsections:

- Patient details
- Journal
- Medication
- Repeat templates
- Drug Sensitivities
- Allergies
- Problem Substances

- Recalls
- Reminders
- Vaccinations
- Cytology
- Summary
- Pathology Results
- Active Problems
- Inactive Problems
- Care Plans
- Social Services Contacts
- Admissions
- Communications

This is how these appear in a sample record for systmOne in Figure 11:

Figure 11 example of a complete systmOne record's overall structure



In this context we are focusing on the three areas of demographic, medical and drug prescriptions the remaining subsections contained repetitive information and therefore no important information was not captured using this approach. For example the journal section contained information about vaccination which was then repeated in the section on vaccination. These subsections are identified and divided by applying the following rules according to the primary regexes for systmOne:

- Look for the word 'Patient' followed by a single space and the word 'Record' and a carriage return. Assign a unique numerical value for this point once it has been reached and remember this value.
- Look for the word 'Journal' followed by a single carriage return. Assign a unique numerical value for this point once it has been reached and remember this value.
- Look for the word 'Medication' followed by a single carriage return. Assign a unique numerical value for this point once it has been reached and remember this value.
- Look for the word 'Repeat' followed by a single space and the word 'Templates' and a carriage return. Assign a unique numerical value for this point once it has been reached and remember this value.
- Look for the word 'Drug' followed by a single space and the word 'Sensitivities' and a carriage return. Assign a unique numerical value for this point once it has been reached and remember this value.
- Look for the word 'Allergies' followed by a single carriage return. Assign a unique numerical value for this point once it has been reached and remember this value.

- Look for the word ‘Problem’ followed by a single space and the word ‘Substances’ and a carriage return. Assign a unique numerical value for this point once it has been reached and remember this value.
- Look for the word ‘Recalls’ followed by a single carriage return. Assign a unique numerical value for this point once it has been reached and remember this value.
- Look for the word ‘Reminders’ followed by a single carriage return. Assign a unique numerical value for this point once it has been reached and remember this value.
- Look for the word ‘Vaccinations’ followed by a single carriage return. Assign a unique numerical value for this point once it has been reached and remember this value.
- Look for the word ‘Cytology’ followed by a single carriage return. Assign a unique numerical value for this point once it has been reached and remember this value.
- Look for the word ‘Summary’ followed by a single carriage return. Assign a unique numerical value for this point once it has been reached and remember this value.
- Look for the word ‘Pathology’ followed by a single space and the word ‘Results’ and a carriage return. Assign a unique numerical value for this point once it has been reached and remember this value.
- Look for the word ‘Active’ followed by a single space and the word ‘Problem’ or ‘Problems’ with or without a colon character ":" and a carriage return. Assign a unique numerical value for this point once it has been reached and remember this value.

- Look for the word ‘Inactive’ followed by a single space and the word ‘Problem’ or ‘Problems’ with or without a colon character “:” and a carriage return. Assign a unique numerical value for this point once it has been reached and remember this value.

- Look for the word ‘Care’ followed by a single space and the word ‘Plans’ and a carriage return. Assign a unique numerical value for this point once it has been reached and remember this value.

- Look for the word ‘Social’ followed by a single space and the word ‘Services’ followed by a single space and the word ‘Contacts’ and a carriage return. Assign a unique numerical value for this point once it has been reached and remember this value.

- Look for the word ‘Admissions ’ and assign and unique numerical value once it has been encountered and remember this value.

- Look for the word ‘Communications ’ and assign and unique numerical value once it has been encountered and remember this value.

These rules can be expressed as the following regexes:

```
@"(\rPatient\sRecord\r)";
@"(\rJournal\r)";
@"(\rMedication\r)";
@"(\rRepeat\sTemplates\r)";
@"(\rDrug\sSensitivities\r)";
@"(\rAllergies\r)";
@"(\rProblem\sSubstances\r)";
@"(\rRecalls\r)";
@"(\rReminders\r)";
@"(\rVaccinations\r)";
```

@"(\rCytology\r)";

@"(\rSummary\r)";

@"(\rPathology\sResults\r)";

@"(\rActive\sProblem(s|:)(.)*\r)";

@"(\rInactive\sProblem(s|:)*\r)";

@"(\rCare\sPlans\r)";

@"(\rSocial\sServices\sContacts\r)";

@"(\rAdmissions\r)";

@"(\rCommunications\r)";

Rationale

- Once numerical values have been assigned to the headings identified in an systmOne record and numerical values have been assigned to the start of the patterns above then it becomes possible to use these numerical values to find the difference in numerical value from one another to ensure a consistent an accurate subsection extract for each regex above . This subtraction defines and creates the start and end of the subsections.

- It is better to use a numerical value to identify start of the subsections so that there is no possibility that there could another set of words that could be confused later on in the whole record. For example there won't be the words 'Medication' causing a conflict between present and past medication as a numerical value is indicated for this regex and it is this value rather than the words that are used in the subtraction that causes the splitting off of these subsections.

- The regexes will mark out the subsection headings using the actual naming convention in the complete record itself. By using case sensitive

notation and combination of carriage returns ensures a highly specific and accurate extraction of subsections into their own text files. It allows the breakdown of whole record into smaller pieces using the natural structure in the record.

4.3.7 systmOne demographic file

Once the demographic subsection is extracted as its own file it is further analysed and broken down by the extraction software to apply the secondary regexes for identifying and cutting out specific demographic related data and inserting this into a new file and thereby creating the demographic file for systmOne. This will produce a systmOne demographic file which contains the following information:

- Name
- NHS Id
- Date of birth
- Gender
- Place of birth
- Marital status
- Ethnic origin
- English speaker
- Address
- Registration date
- Date of removal
- PDS registered practice
- Registered GP
- Registration type
- Medical records
- Dispensing regex
- Pharmacy

- Appointments
- Attendance
- Visits
- Did not attend (DNAs)
- Practice Id
- Deceased date

The remainder of this section will describe show how this occurs, fictional samples and rationale to support the design of the regular expressions applied.

4.3.7.1 systmOne: data extraction of the Patient name, NHS Id and Date of birth

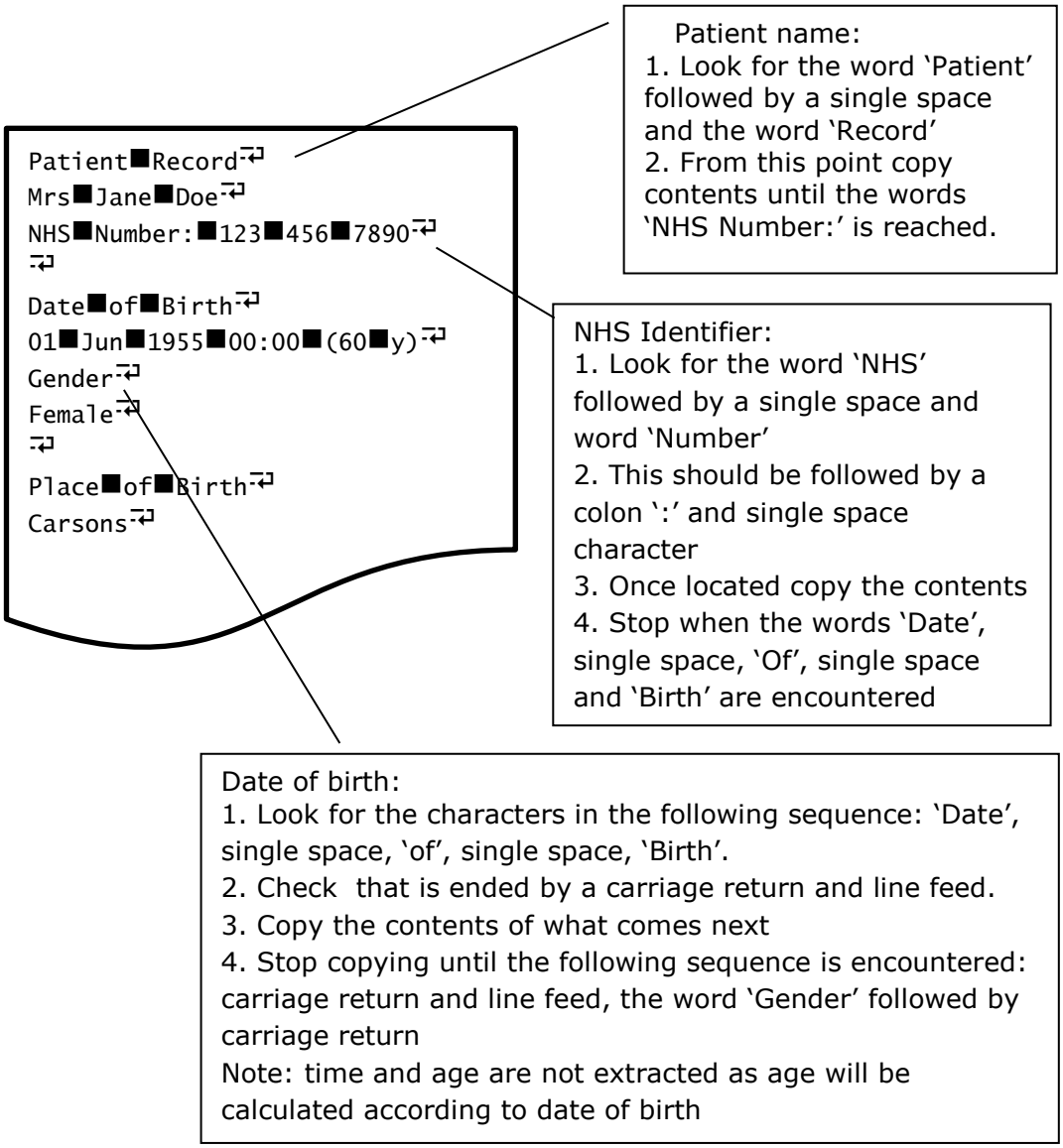


Figure 12 – systmOne: extraction of name, nhsid & date of birth
Key:
↵ carriage return & new line
■ space
⇊ tab

There is no specific labelling of the name variable in systmOne records; however the variable is identifiable in the context what occurs around it.

4.3.7.2 systmOne: Patient Name

There is no specific labelling of the name variable in SystmOne records, however the variable is identifiable in the context what occurs around it.

Patient name is extracted using the following regular expression:

@"(\patient\record\r)(?<name>(.)*)(\rNHS\sNumber:\s)";

Rationale:

- The patient name variable is not labelled explicitly.
- It is always found between the words listed above.
- This method allows capture of whole name.

4.3.7.3 systmOne: The NHS Identifier

The unique NHS Id is labelled and appears consistently in the same place for each patient. For systmOne records it has unique pattern of spaces as part of the 10 digit number which is then completed by a carriage return at the end of the line.

NHS Id is extracted using the following regular expression:

@"(\rNHS\sNumber:\s)(?<nhsId>(.)*)(\r\rDate\sof\sBirth\r)";

Rationale:

- The record has its own label associated with the data item which can easily locate the item.
- It's better to copy after the words in step 1 (from the above 'Patient' box in Figure 12) as it ensures only the data item and not its label are being extracted.

- In case there is a problem with lack of spaces this method still captures the data item properly as it extracts all the data between the words indicated, NHS Id and Date of Birth labels.

4.3.7.4 systmOne: Date of Birth

This data item has its own label associated with its appearance in the record and occurs in the same position each item for all patient records.

Date of birth is extracted using the following regular expression:

@"(\r\rDate\sof\sBirth\r)(?<dob>(.*))(\rGender\r)";

Rationale:

- It uses the same sequence of capitalisation and spaces to enable the label to identified accurately. By copying between two labels of 'Date of Birth' to 'Gender' ensures that the whole of date of birth is copied irrespective of which date format is used.
- If the vendor decided to change the data format from dd/mm/yyyy to something else this regex extract would still work.

4.3.7.5 systmOne: Extraction of Gender, Place of Birth & Marital Status

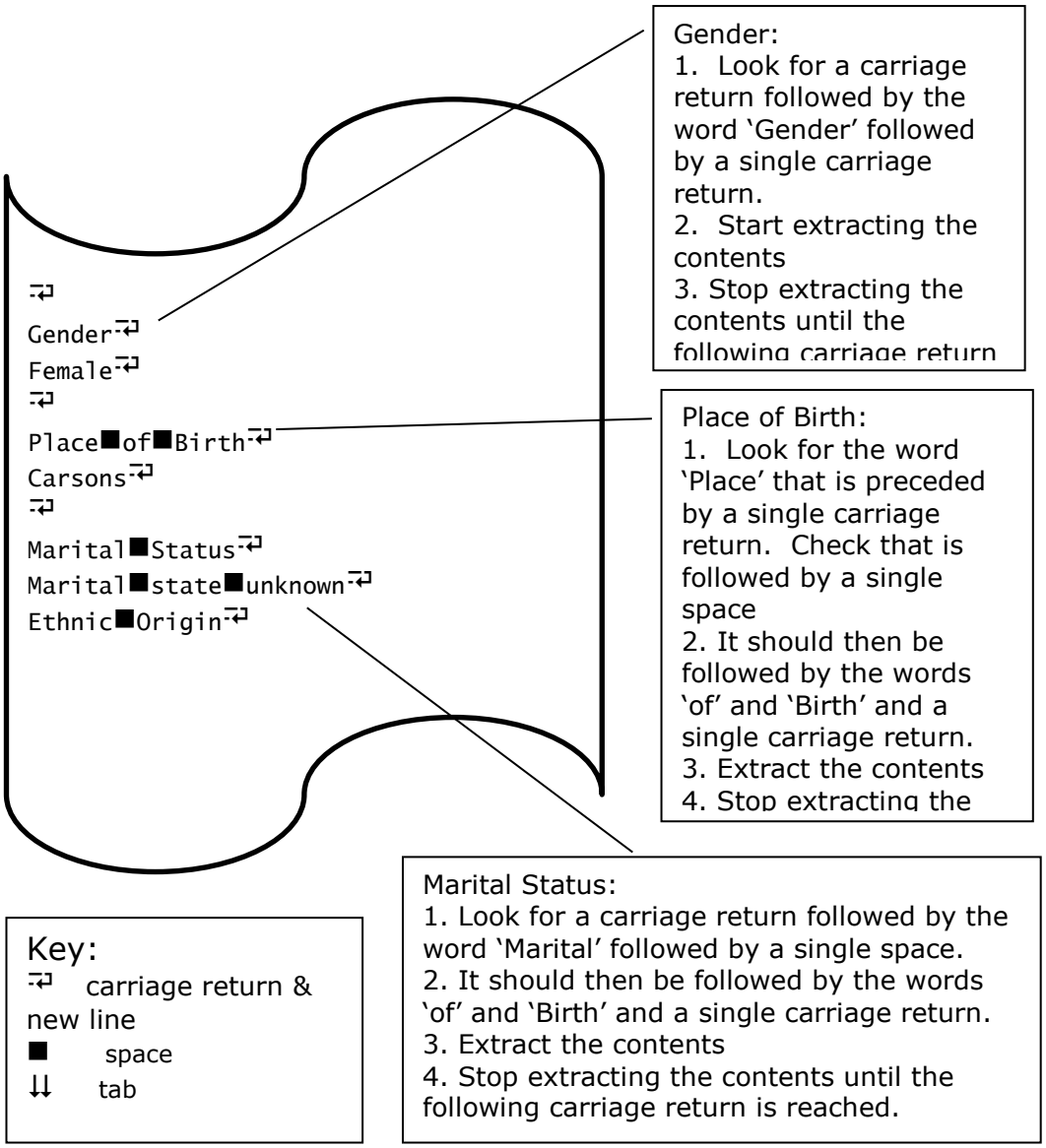


Figure 13 - systmOne: extraction of Gender, Place of Birth and Marital status

4.3.7.6 systmOne: Gender

This data item corresponds accurately to its relative label so can be identified and extracted using its own unique label. It comprises of the patient's gender completed by a carriage return. Gender is extracted using the following regular expression:

@"(\rGender\r)(?<gender>(.)*)(\r\rPlace\sof\sBirth\r)";

Rationale:

- This item has its own label identifying its location.
- There does appear to be an empty line with only a carriage return once the gender value is displayed. If in the future anything appears on this line or the user enters some free text on this empty line between the next label the gender value will still be extracted even if a little variation occurs around the value itself.

4.3.7.7 systmOne: Place of Birth

This data item does have its position indicated by its own unique label and can be distinguished from anything similar such as address.

Place of Birth is extracted using the following regular expression:

@"(\rPlace\sof\sBirth\r)(?<placeofbirth>(.)*)(\r\rMarital\sStatus\r)";

Rationale:

- Place of birth is a very ambiguous label however it is fixed. If considered in the context of the variety and format that this data could be entered, for example place maybe a county, city, or country.
- As a result it is better design to capture the contents between the two labels in case there is variation across practices about how place of birth

is defined and what information is placed here within this placeholder for birthplace.

4.3.7.8 systmOne: Marital Status

The marital status data item is indicated with its own label within the record and this is used to identify it in a unique and consistent manner. The following rules apply when locating the marital status:

Marital status is extracted using the following regular expression:

@"(\rMarital\s\Status\r)(?<maritalStatus>(.)*)(\rEthnic\sOrigin\r)";

Rationale:

- Marital status can have different choices in potential data items such as married, divorced, single, remarried or separated.
- The definition list may be updated to changes by GP software systems or even in legal terms; as a result it is much more robust to extract the contents that occur between the two labels.

4.3.7.9 systmOne: Extraction of Ethnic Origin & English Speaker & Address details

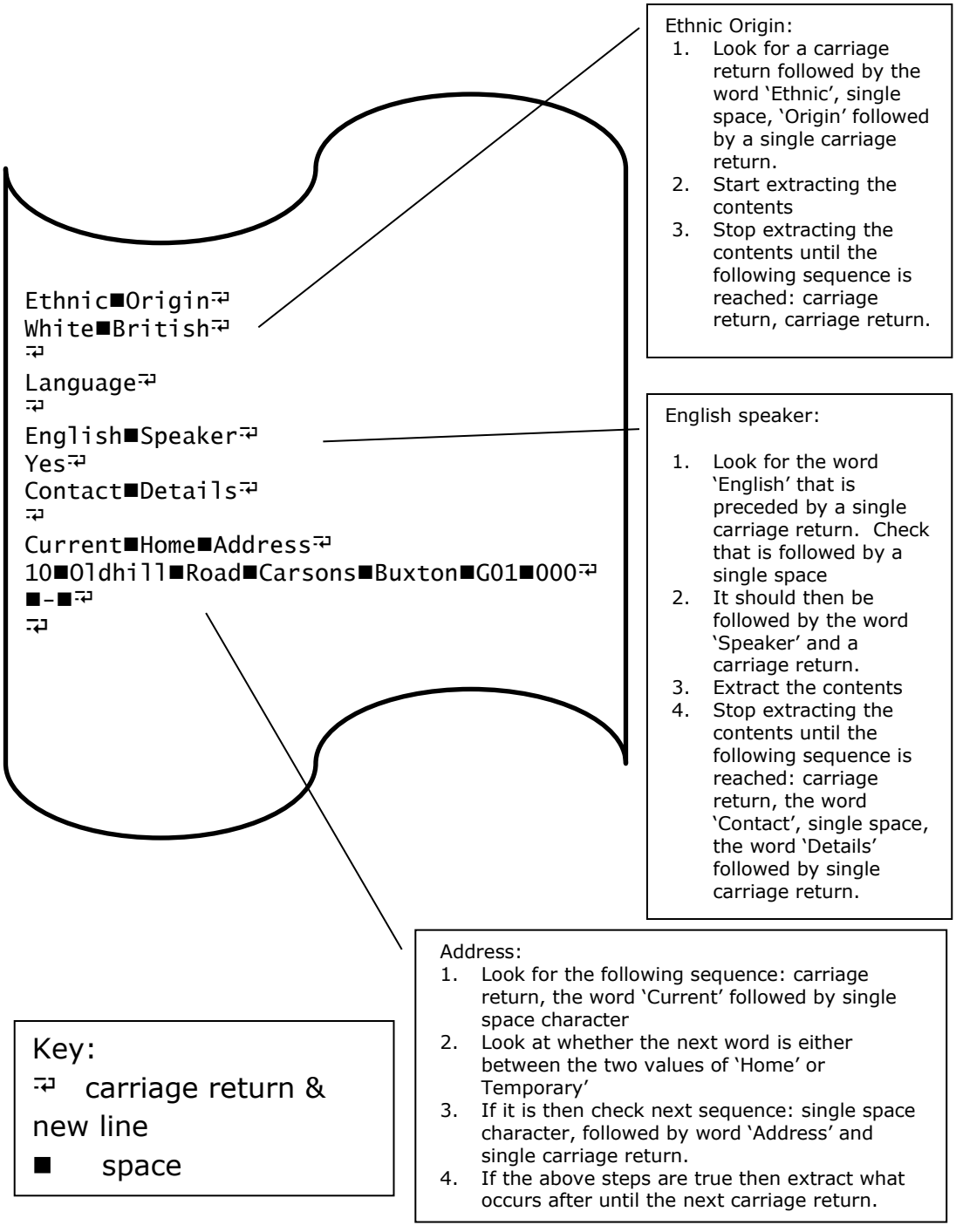


Figure 14 - systmone: extraction ethnicity, language and address

4.3.7.10 systmOne: Ethnic Origin

Ethnic origin has its own unique label attached to it to indicate where it is to be found and therefore can be distinguished from other data items.

Ethnic origin is extracted using the following regular expression:

@"(\rEthnic\sOrigin\r)(?<ethnicOrigin>(.)*)(\r\rLanguage\r)";

Rationale:

- For reliability it is much more robust to extract ethnicity value from between the two data labels than create a list from looking at the current values found in the records to allow for differences

4.3.7.11 systmOne: English Speaker

This data item indicates whether the person speaks the English language as indicated by an answer of yes or no. It also has its own label attached to the data item indicating its place in the record.

Rationale:

English speaker is extracted using the following regular expression:

@"(\rEnglish\sSpeaker\r)(?<engSpeaker>(.)*)(\rContact\sDetails\r)"Rationale:

- This allows for a variety answer beyond yes/no including 'speaks a little' or 'does not understand at all'.

4.3.7.12 systmOne: Address

This refers to capturing the current address of the patients. In SystmOne there are two ways of address definition: it can be either current home or current temporary address.

Address is extracted using the following regular expression:

@"(?<=(\rCurrent\s(Home|Temporary)\sAddress\r))(?<address>(.)*)\r";

Rationale:

- Address has a variable way that it begins as its label can take on more than option. As a result the regex needs to be able to handle the options available.
- There is no other consistent label that can be used as an end marker as in previous data items above as what follows in the mobile number which not everyone has on their record.
- The full address end of line marker using a carriage return is consistent across all records to allow accurate extraction.

4.3.7.13systmOne: Extraction of registration date & date of removal

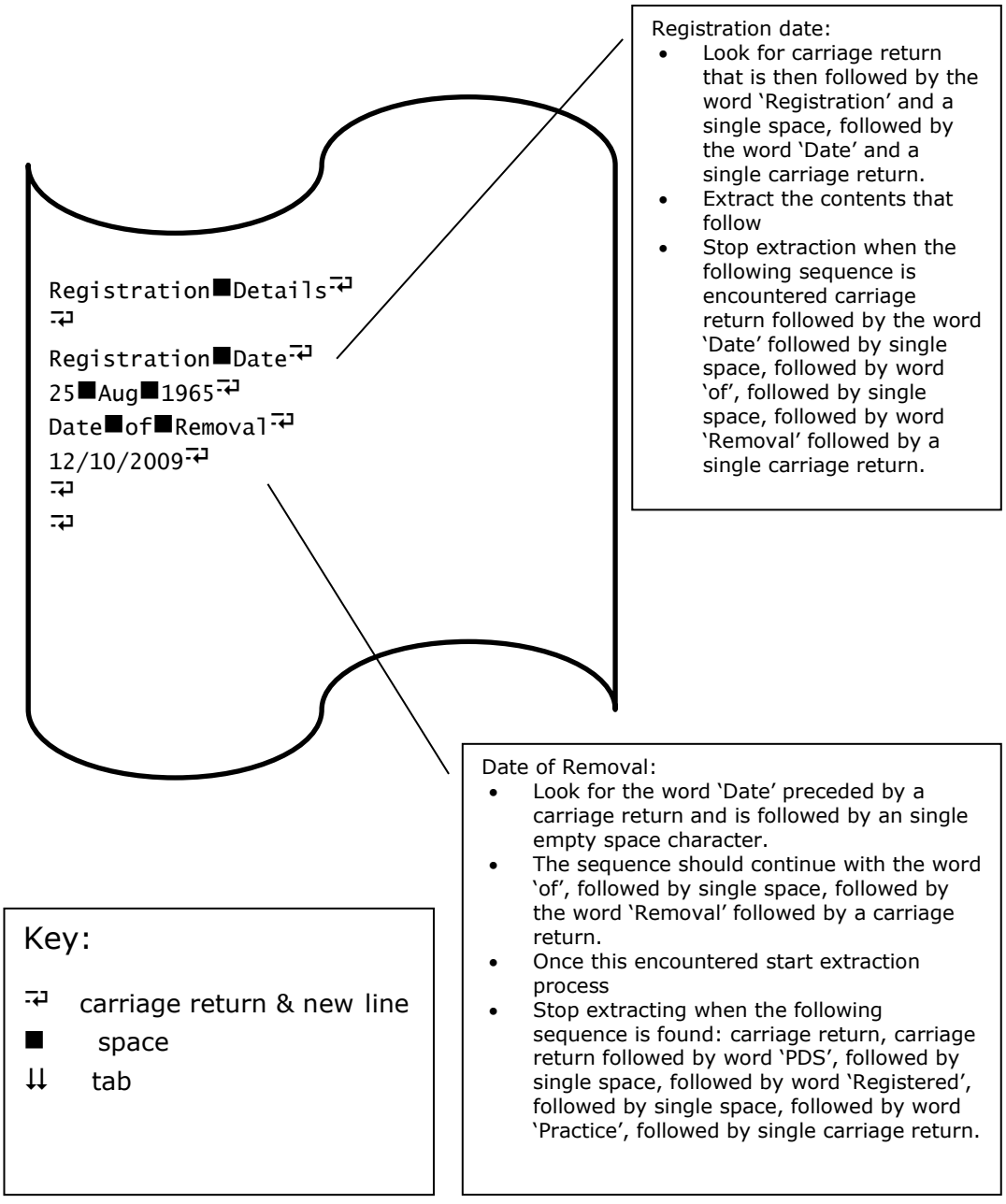


Figure 15 – systmOne: extraction of Registration Date

This refers to the date that patient registered at the practice as a patient. It can be entered in a variety of ways for example as dd/mm/yyyy or sometimes it can only be the month and the year concerned. Registration date is extracted using the following regular expression:

@"(\rRegistration\sDate\r)(?<registrationDate>(.)*)(\rDate\sof\sRemoval\r)";

Rationale:

- Registration date's format can take on various standards.
- There may be added bits of information as to anything that supplement the registration date that could be added in by the practice itself.

4.3.7.14 systmOne: Date of Removal

This indicates the date at which the patient was removed from the practice. The date of removal can be for either death of patient or date that a patient has moved away. If the date is empty no data will be extracted and a character space will be present in the extracted file. The following rules apply when trying to locate this date:

Date of removal is extracted using the following regular expression::

@"(\rDate\sof\sRemoval\r)(?<dateOfRemoval>(.)*)(\r\rPDS\sRegistered\sPracti
ce\r)"

Rationale:

- Not all patients will have a date of removal as they still current patients
- As a result many times this value will not be present and rather than explicitly look for a date structured value it is better to extract the contents.

Whether or not there is anything actually present as content it will still return whatever is found between the two variables and this may not only be a date value but other information such as 'moved to another country' for example.

4.3.7.15systmOne: Extraction of Personal Demographic Service (PDS)

Registered Practice, Registered GP and Patient Registration Type

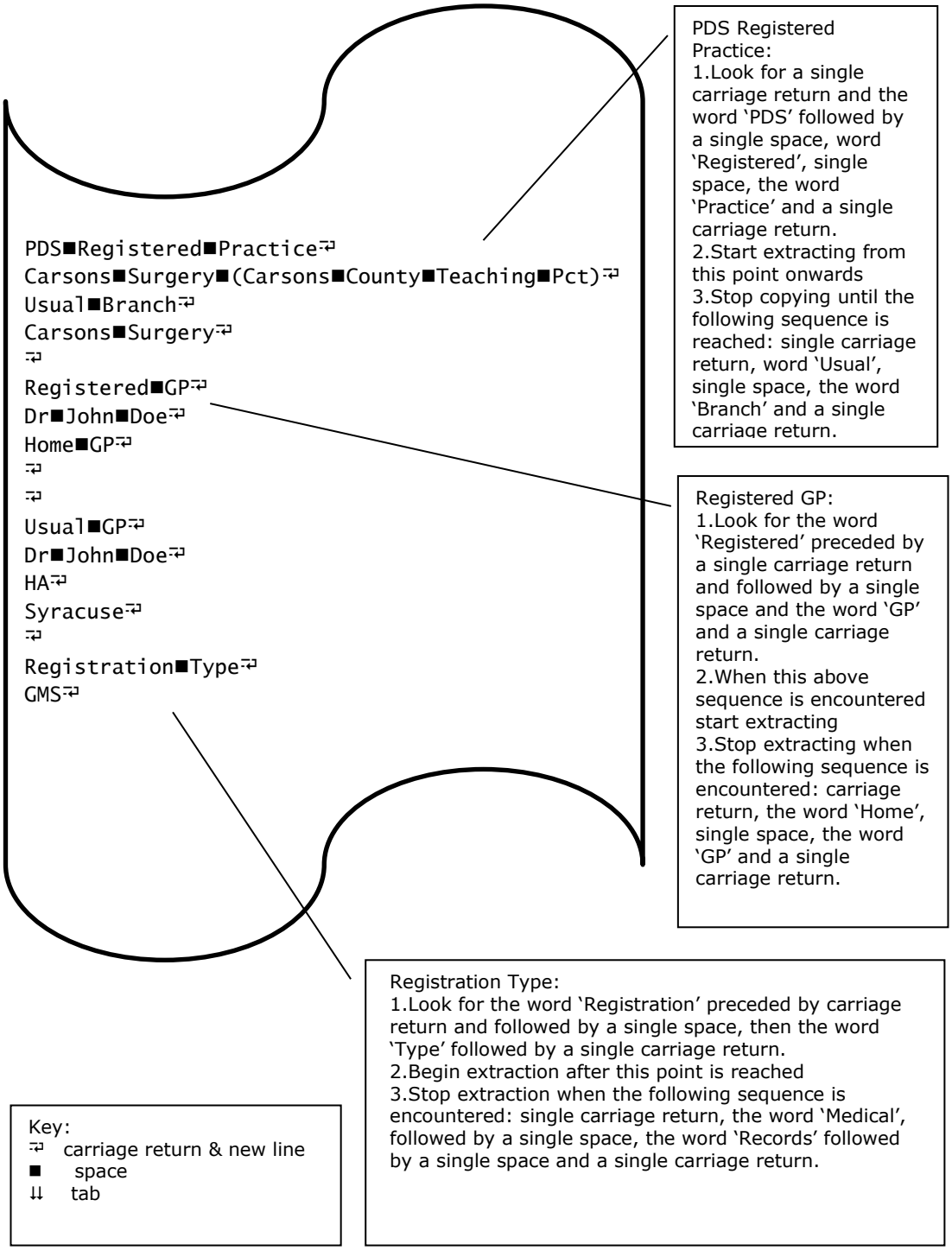


Figure 16 - systmOne: extraction of PDS Registered Practice

4.3.7.16systmOne: Personal Demographic Service Registered Practice

This indicates the practice is Personal Demographics Service (PDS) registered practice. It contains the practice name and medical group if any that it may be attached with.

Personal Demographic Service registered practice is extracted using the following regular expression:

@"(\rPDS\sRegistered\sPractice\r)(?<pdsRegisteredPractice>(.)*)(\rUsual\sBranch\r)";

Rationale:

- This data item contains a variety of information including the practice name and any other practice related details
- As a result it is better to use the convention given in the records rather just practice name as there may be information of use in the this section even though it's not directly related to patient demographic information.

4.3.7.17systmOne: Registered General Practitioner (GP)

This data item refers to the patient's registered GP

Registered GP is extracted using the following regular expression:

@"(\rRegistered GP\r)(?<registeredGp>(.)*)(\rHome\sGP\r)";

Rationale:

- In many records it is actually the name of a medical group rather than the GP that is present

- There are maybe cases where there is medical group information here instead of an actual person and so it is more robust to extract all the data between the two variables.

4.3.7.18 systmOne: Registration Type

This indicates that the registration type of the practice such as General Medical Service (GMS) for example. It refers to the range of services that the GP is registered to provide.

Registration type is extracted using the following regular expression:

@"(\rRegistration\sType\r)(?<registrationType>(.)*)(\rMedical\sRecords\r)";

Rationale:

- The data item extracted is normally an abbreviation of a term for example 'GMS'. It's possible just to look for the abbreviated term but this would not be sensible in case Government policy defines new abbreviations and deprecates older ones.
- There may be additional freetext information attached to the registration type which may be of interest

4.3.7.19systmOne: Extraction of Location of Medical Records, Dispensing & Pharmacy

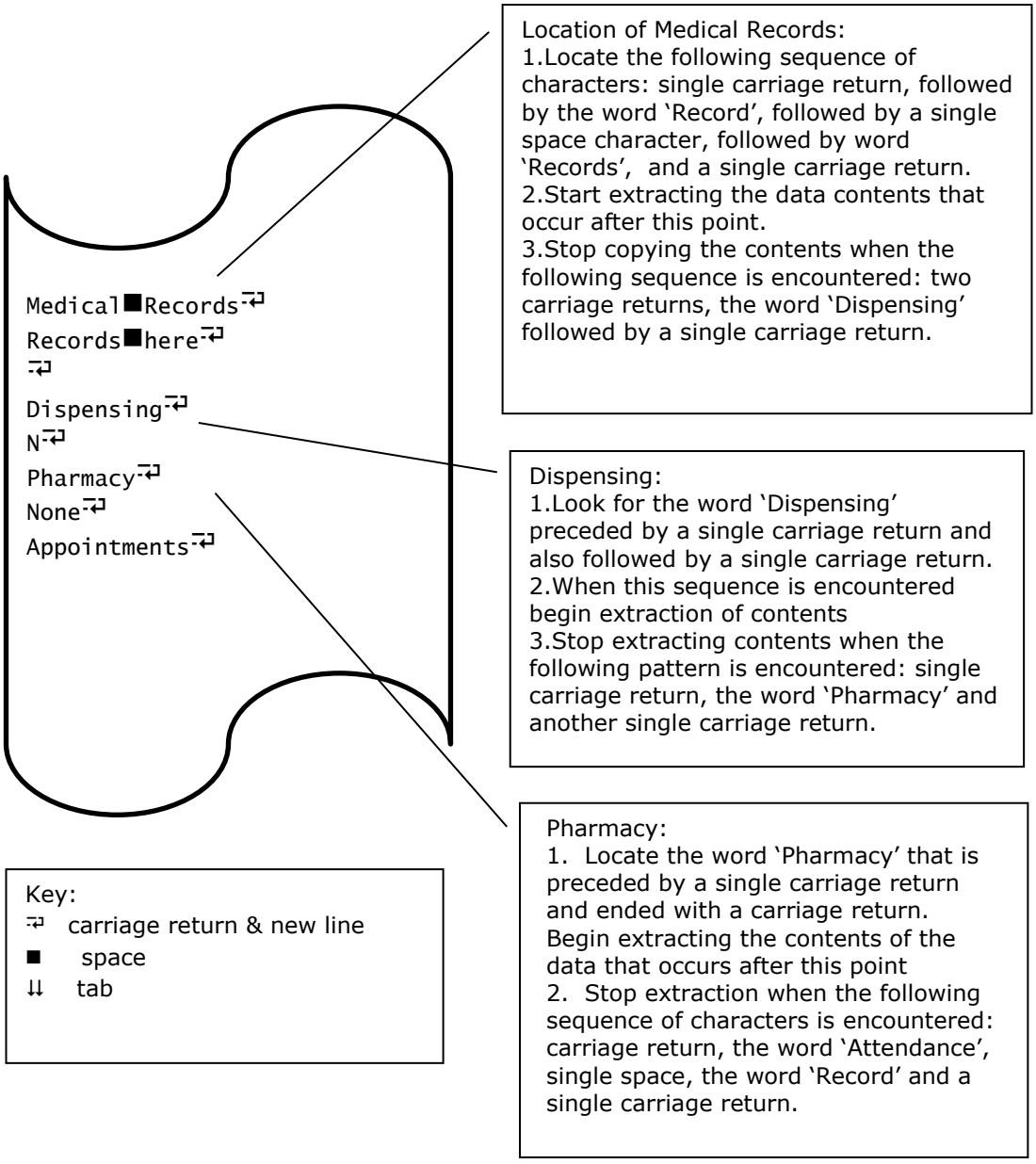


Figure 17 - systmOne: extraction of medical records, dispensing and pharmacy

4.3.7.20 systmOne: Location of Medical Records

This is more linked to practice level variable rather than to the patient details and refers to the location of the medical records for the patient.

Location of medical records is extracted using the following regular expression:

@"(\rMedical\sRecords\r)(?<medicalRecords>(.)*)(\r\rDispensing\r)";

Rationale:

- The outcome can take on a variety of customised definitions depending on practice management structure and storage facility.
- As a result it is more robust and better design to ensure any information is extracted and of interest to extract the contents between the two variables indicated.

4.3.7.21 systmOne: Dispensing

This information is regarding whether the practice is a dispensing one. Although not directly related to the patient demographic details, it may provide interesting information. Dispensing is extracted using the following regular expression:

@"(\rDispensing\r)(?<dispensing>(.)*)(\rPharmacy\r)";

Rationale:

- The data extracted for this item will always either be 'Yes' or 'No', although the possibility of a transitional phase such as where a practice has applied to become a dispensing practice may be indicated here as well.

4.3.7.22 systmOne: Pharmacy

This piece of data item indicates whether there is a pharmacy present at the practice. Pharmacy is extracted using the following regular expression:

@"(\rPharmacy\r)(?<pharmacy>(.)*)(\rAttendance\sRecord\r)";

Rationale:

- Pharmacy outcomes do not have a set and specific list of outcomes such as yes, no, none, and none at present but in process of becoming so in future.

4.3.7.23 systmOne: Extraction of Appointments, Attendance, Visits & Did Not Attends (DNAs)

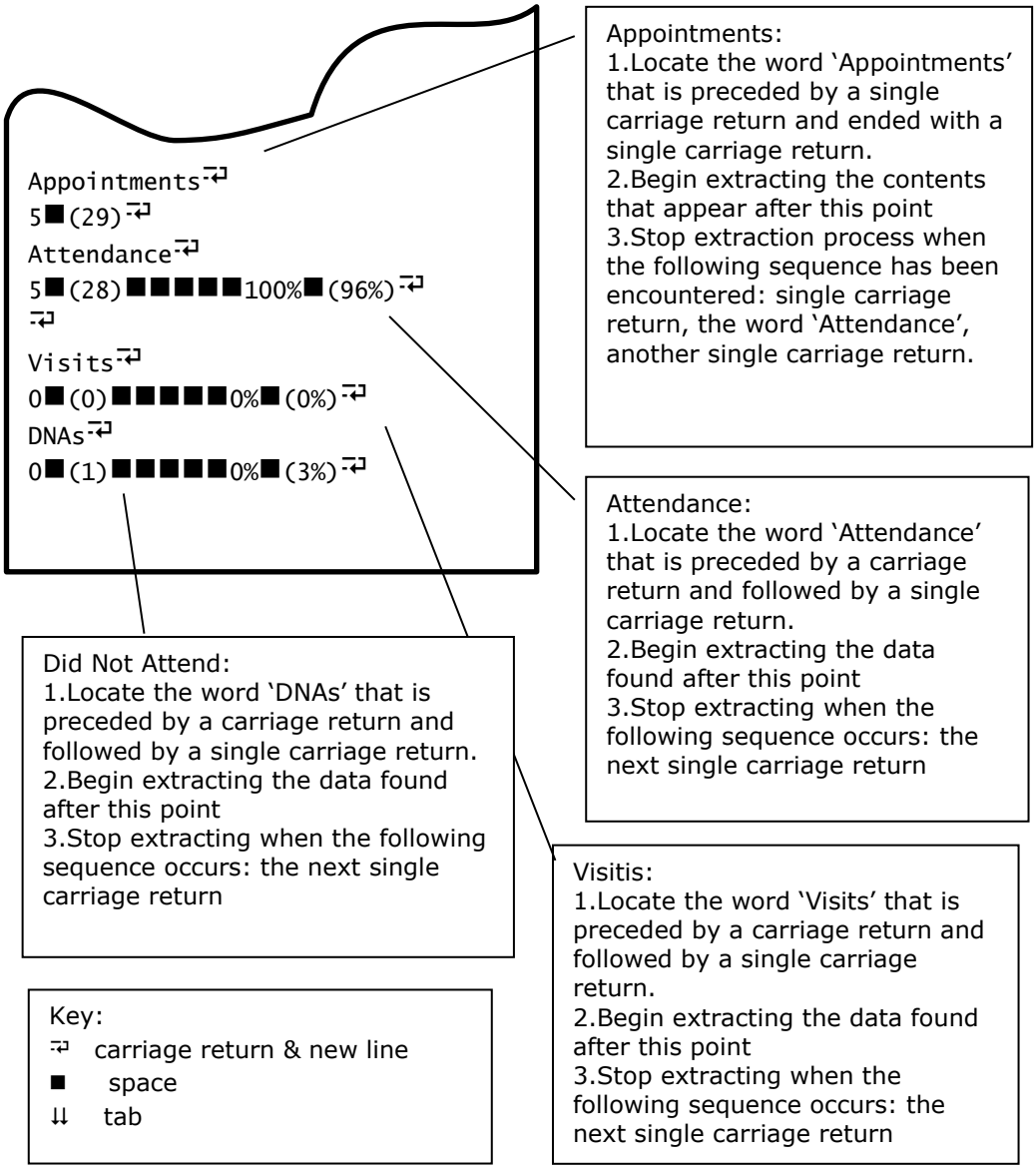


Figure 18 - systmOne: final segment of record concerning the extraction of appointments, attendance, visits and did not attends

4.3.7.24systmOne: Appointments

This refers to the appointments history that the practice keeps of the patient concerned. Appointments are extracted using the following regular expression:

@"(\rAppointments\r)(?<appointments>(.)*)(\rAttendance\r)";

Rationale:

- Appointments data item is formatted as a collection of numbers and open and close brackets which together represent the attendance record of the patient
- It cannot be explicitly coded to look for a specific target of numerical characters as this can change and be more or less for each patient.
- As a result to ensure all appointment data is captured it is better to indicate that all the information between the two labels is extracted.

4.3.7.25systmOne: Attendance

This refers to the attendance history that the practice keeps of the patient concerned. Attendances are extracted using the following regular expression:

@"(\rAttendance\r)(?<attendance>(.)*)(\rVisits\r)"

Rationale:

- The attendance data item is formatted as a collection of numbers and open and close brackets which together represent the attendance record of the patient
- It cannot be explicitly coded to look for a specific target of numerical characters as this can change and be more or less for each patient.
- As a result to ensure all attendance data is captured it is effective to indicate that all the information between the two labels is extracted.

4.3.7.26 systmOne: Visits

This refers to the visits history that the practice keeps of the patient concerned. Visits are extracted using the following regular expression:

@"(\rVisits\r)(?<visits>(.)*)(\rDNAs\r)"

Rationale:

- The visits data item is formatted as a collection of numbers and open and close brackets which together represent the attendance record of the patient
- It cannot be explicitly coded to look for a specific target of numerical characters as this can change and be more or less for each patient.
- As a result to ensure all visits data is captured it is better to indicate that all the information between the two labels is extracted.

4.3.7.27 systmOne: Did Not Attend

This refers to the 'Did not attend' (DNAs) history that the practice keeps of the patient concerned. DNAs are extracted using the following regular expression:

@"(\rDNAs\r)(?<dnas>(.)*)(\r)

Rationale:

- The DNAs data item is formatted as a collection of numbers and open and close brackets which together represent the attendance record of the patient
- It cannot be explicitly coded to look for a specific target of numerical characters as this can change and be more or less for each patient.
- As a result to ensure all DNAs are captured it is better to indicate that all the information between the two labels is extracted.

4.3.8 Extraction of Medical file data for systmOne

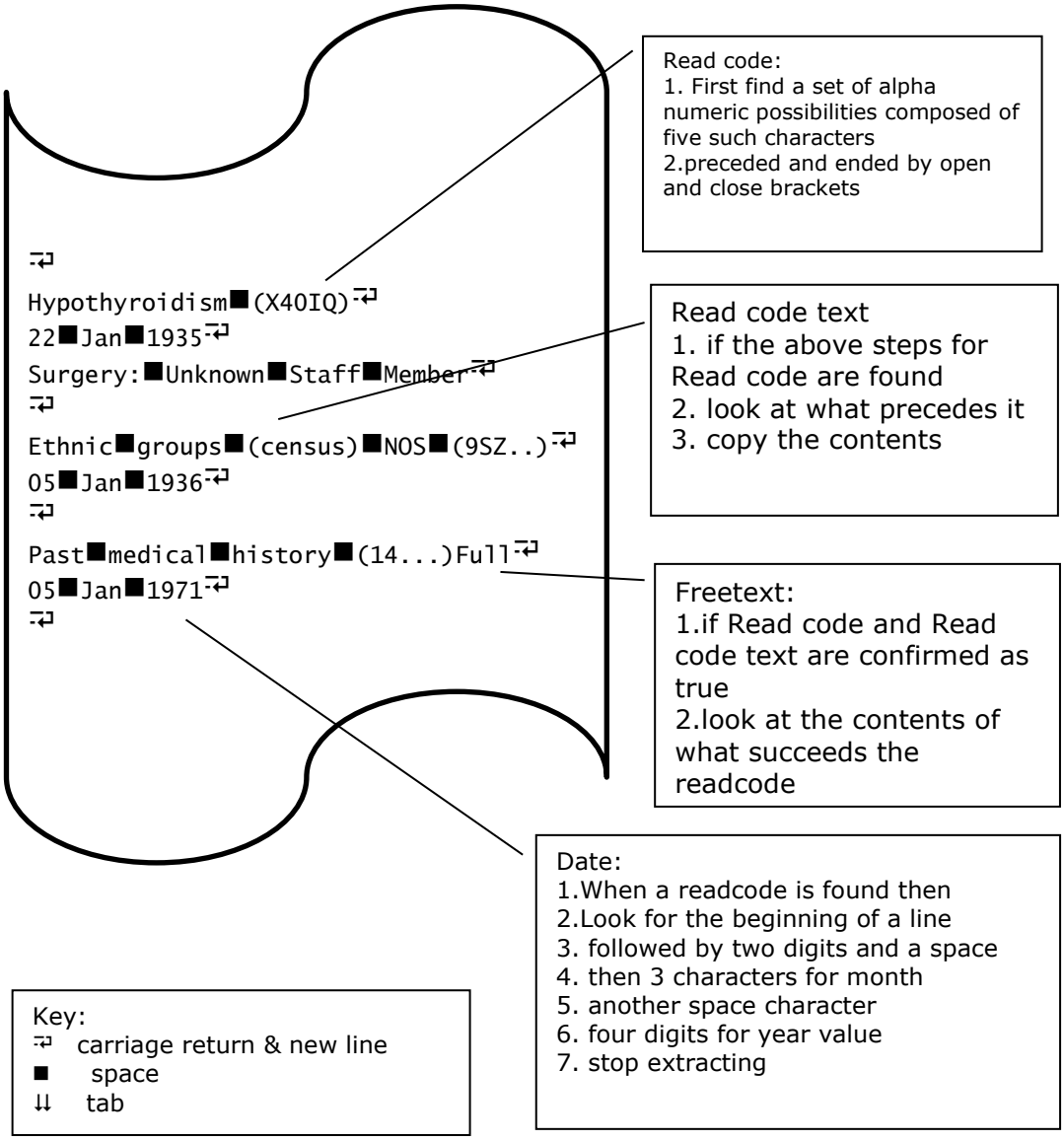


Figure 19 - systmOne: medical extraction date, Read code, Read code text and free text

4.3.8.1.1 systmOne: Read code, Read code text, free text and date

This refers to the therapy that the patient has received. In systmOne records all the therapy prescribed is encoded in terms of Read codes. This means that for every patient the following is extracted:

- Read code Text – this refers to the name given to the Read code itself. It is a description in English of what the Read code represents.

- Read code – This is the Read code itself and represents a 5 digit unique identifier of the therapy prescribed.
- Read code Free Text – This refers to the supporting text that follows the Read code itself and is ad-hoc text inserted by the doctor or practice themselves
- Date – Each therapy received has a date associated along with it; so it becomes possible to define a pattern that matches the date the patient received the therapy indicated by the Read code.

When looking at the EMR section that displays the therapy it is possible to design a regex pattern that extracts the above pieces of information of Read code text, Read codes and any free text using a 'chained' regular expression:

(?<readcodeText>(.*?(?=\([A-Za-z0-9|\.]{5}\)))(?<readcode >\([A-Za-z0-9|\.]{5}\))(?<readcodeFreeText>(.*))

Rationale:

- The parts above always occur on the same line and never in any other way, so it is effective to write the regex in the same manner so it is easily distinguishable from other lines that are present within the record.
- Coupling several regexes together this ways helps in locating the bits of data being sought by ensuring a very specific pattern is used
- Once the chained regex is found to be true then the medical date is extracted using the following regular expression:

@"(?<date>^(\d{1,2}\s\w{3}\s\d{4}))(\s\d{2}\:\d{2})?$"

Rationale:

- This approach of indicating that only when a date value begins a line should it be considered takes advantage of how this is structured in the actual record layout of systmOne.

4.3.8.2 Checking of Read or Clinical Codes

An extra step of validation of extracted Read codes was used. This was achieved by matching Read codes against a dictionary list of terms provided by the Health and Social Care Information Centre (HSCIC) so that the corresponding Read codes could be checked as being valid. At the time of writing the most recent, comprehensive and officially used clinical codes for the United Kingdom were downloaded using the Technology Reference data Update Distribution (TRUD) site which was owned and run by the Health and Social Care Information Centre (HSCIC) and since July 2016 this has been rebranded as NHS digital. The pack downloaded was used for searching for both Read version 2 and Clinical Terms Version 3 (CTV3) was the 15.0.0 release pack and covers over 150,000 clinical codes (141). (See appendix 8.10 for more information). If a Read code failed to be validated then the code was instructed to move the description to free text and supply a message on an error file produced in the EMR source folder. Post software and data testing of all systmOne files no errors were reported in the medical Read code file.

4.3.8.3 systmOne: Capturing Free Text in Medical Data

Additional free text may include all kinds of useful bits such as blood pressure readings, results, extra comments and anything related to the patient therapy. As a result if a line is encountered that does not match the above it is placed automatically in the free text section of the data extracted, that can be processed in the future.

4.3.9 Extraction of Drug Prescriptions for systmOne

Figure 20 shows the structure of the drug file prescriptions with 2 fictional prescriptions in it.

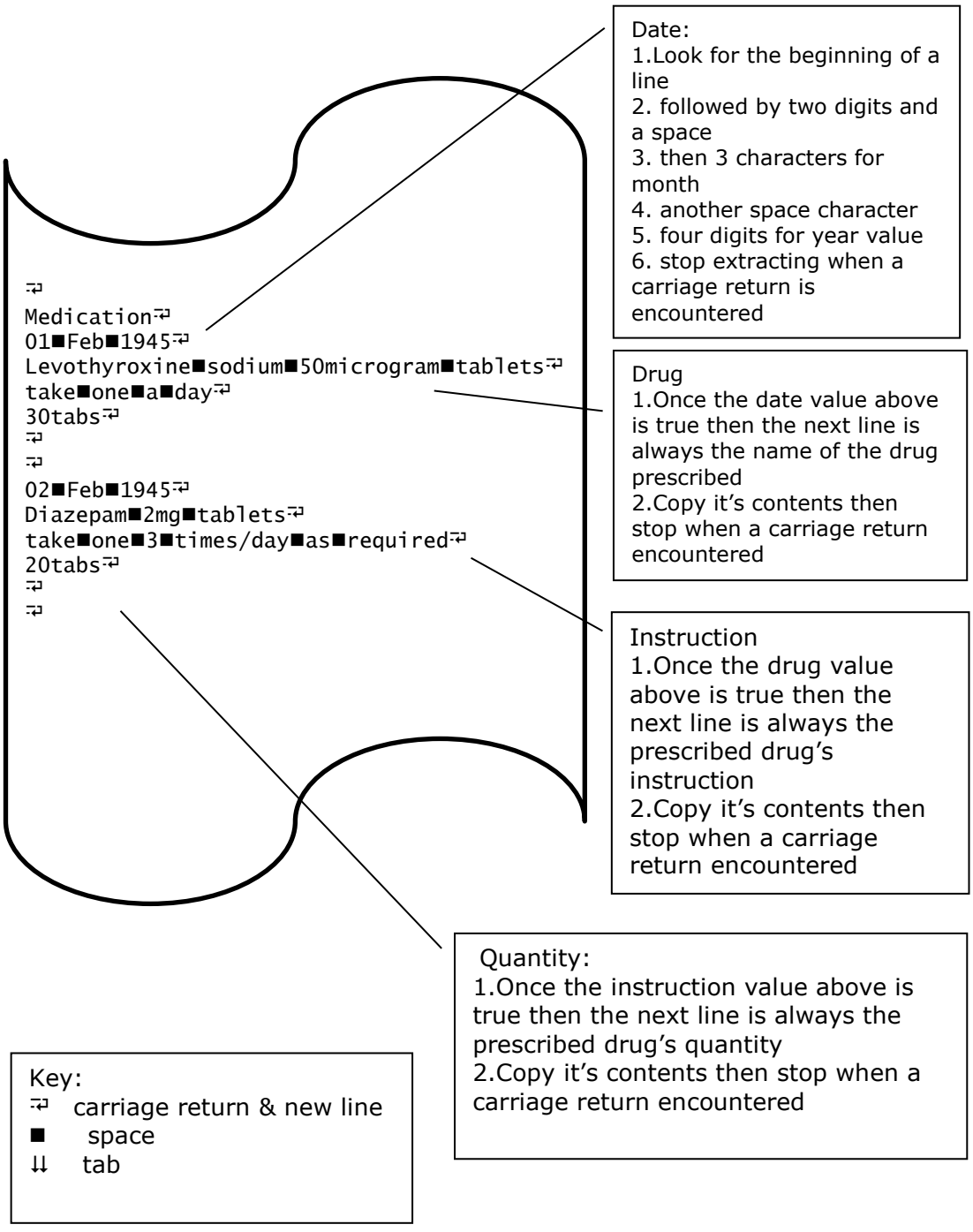


Figure 20 – systmOne: drug extraction of date, drug name, dose, form, instruction and quantity

4.3.9.1.1 systmOne prescriptions: Date, Drug, Instruction and Quantity

Prescriptions in SystmOne are structured in a very specific way over a specific number of lines for each prescription. This has enabled a coupling of regexes to occur to enable the unique structure to be looked at as a whole rather than several lines. The following rules apply when looking for the drugs prescribed for a patient:

- Date of Drug Prescription – Date that the drug was prescribed to the patient.
- Drug Prescribed – Actual drug itself prescribed by the physician.
- Instruction – general instruction for the course duration.
- Quantity – Number of tablets for example.

These coupled regexes are expressed in the single main expression below:

```
@"(?<drugDate>(\d{1,2}\s\w{3}\s\d{4}))(\r)

(?<drug>([^\r])*)(\r)

(?<instruction>([^\r])*)(\r)

(?<qty>([^\r])*)(\r)";
```

Rationale

- All four key parts of: date, drug, instruction, quantity occur in the same pattern for each entry of drug prescribed.
- This presents an opportunity to couple several regexes together in a single expression. It also ensures greater specific application of the regex to extract only the regex patterns contained within the single main expression.
- This allows a line by line extract for each entry in the journal so that a patient's drug instruction from one entry would get mixed up with the

instruction from another drug prescription on the same or any other entry.

4.4 Producing the iSOFT Data Files

4.4.1 Introduction to iSOFT data extraction

The iSOFT GP system is produced by Computer Sciences Corporation (CSC).

4.4.2 Downloading the iSOFT files

An electronic PDF (portable document format) version of the iSOFT EMRs were taken and then processed using the extraction software (ES) into three usable files which covered the data concerned demographics of the patients, medical records and through to the drugs prescribed by a series of regular expressions. See appendix 8.9 for systmOne EMRs download instruction.

4.4.3 PDF-text file conversion

The output file is a .pdf however Microsoft Corporation's current release of their software framework (for C#) does not include a way that the computer can view pdf-text in the same way as it can 'see' standard text files such as the ones in a notepad file for example, therefore the first step that is needed is to convert the .PDf into a text file. Without the support inside the programming language an intermediary process to do the conversion is required. The options explored are below but selecting the best approach was based on the following needs it must fulfil:

- Low cost or Freely available
- Conversion process must be 'true to life' so that the copied text closely matches the original contents structure as close as possible
- Consistency across all records
- Least time consuming and easy on the user as possible

Having identified the need for a conversion step from .pdf to .txt (text file) several avenues were explored:

4.4.3.1 Manual Conversion Options:

Option 1: Notepad Plus - The first was to create a new document in a software called Notepad Plus. This was done by opening the file in inside the free to use

pdf viewer called Adobe Reader XI (from Adobe Systems, Inc.) and copying the lines from the PDF file and then pasting the copied text into Notepad Plus, after which the option 'Encode in UTF-8 without BOM' (byte order mark) (in the Encoding menu bar) was selected. The window was then saved as a normal text file from the 'file' > 'save as' option. This was attempted with the view to minimise the loss in structure and visibility between characters from .pdf and text. However this option was rejected as it did not render the structure similar to the original pdf. This is important since if it is not possible to obtain original text file versions than any converted EMRs should match the originals as closely as possible so that if in future it's possible to obtain original text files from iSOFT then any changes to the regexes will be very little.

Option 2: Adobe Reader XI and use 'save as text' - A second option was to use the facilities in Adobe Reader XI. This was done using the option 'file' > 'save as text' option and a file is then created with the text version of the PDF file. This was rejected due to the process interfering too much with the original structure.

Option 3: Adobe Reader XI and Copy All Contents - Another option available with Adobe Reader XI is its ability to copy all the contents of its current file to the 'clipboard' memory of the computer. This was done using the option 'edit' > 'copy file to clipboard' option. Once this was selected a new Notepad document was opened and the option 'edit' > 'paste' was selected. All the PDF contents then become immediately available in the Notepad window and the 'file'>'save as' option can be used to save the text file. Again this was not a consistent conversion and it would be impractical to do this for every single PDF page of the original EMR.

Auto-Processed PDF to Text Using Software

Option 4: PDF Converter 7 - A further option explored was to use an 'off the shelf' software package to do the conversion process from PDF to text

characters. The software used was called 'PDF Converter 7' from Nuance Communications, Inc. This worked by the user selecting the folder where the pdf files are stored then choosing the destination where new ones would be after conversion and then clicking the process button to do the conversion. Text file versions of the PDF files then become available in the destination folder. Although this worked consistently across all records and it was quick to convert them all a problem occurs because it's using a third party commercial product. It is inappropriate to lock down the Extraction Software (ES) to a third party software as it means any changes in the company's products which is opaque to the created ES can affect the working and hence accuracy of the data files extracted. It is also an added expense as it was necessary to purchase the product before it could be tested.

Option 5: Open Source Software Translator – Poppler Utils is an open source software application. It has a library that can translate or convert from different types of files to a variety of others and in this case it's the pdftohtml.exe 'library' that was tested to enable a conversion of PDF files to XML (eXtensible Mark-up Language). Incorporating Poppler Utils allows the user to feel that no extra step is necessary during the conversion process as it occurs in the background. The testing showed a very stable conversion that was accurate, kept to the original EMR structure better than the above options thus far and is also free to use and incorporate into the software without any financial costs.

4.4.4 Choice of PDF Conversion Tool

The options were weighed within this context; using option 1 is not practical as it is a time consuming process for the user to copy and paste the record data. Though it is free and 'true to life' and is consistent it was rejected due to the unreasonable burden on the users. This serious drawback also affects options 2 and 3 using Adobe Reader XI. Another problem with Adobe Reader XI was that in many places the text was not converted in a consistent manner with the

original EMR with lines containing carriage returns that were not in the original. PDF converter 7 was considered next; the software costs around £60.00 pounds to purchase and it converts PDF documents in a consistent and an accurate way but suffers from the drawback that using specific software means the ES is locked down to the specific PDF software concerned. Being linked this way to one commercial software system is not a sensible strategy in long term as the ES is then reliant on that. It is best to develop the ES in a manner that is independent of other software interpretations. There was also too much burden on the user with this process as it is manual conversion process.

Lastly Poppler Utils, this software is freely available under a General Public Licence (GPL) which means it is free to use and distribute. It fulfilled all aspects of the above criteria because it converted text in way that required no extra input from the user and is free to use. It is called from within the data centre itself and so becomes a part and parcel of the data centre. To the user this would seamless that the conversion and extraction has all been done at the same time. It also converted the PDF in a way that allowed the regex engine to interpret the converted text accurately without losing any data and this conversion process was checked to ensure that it was consistent both within the record and with all other records. As a result it was the chosen method to convert the PDF files into a text format that could be used by the data centre for the regex extractions to occur. The Poppler Utils software converts the PDFs into an .xml or XML file that is then used to cut the record into its subsections. The XML specific structure is examined (this structure is a text file format from the regular expression's perspective) and the headings are marked out and then regex's developed according to these. The Calibre library is open source under the terms of the GNU Public Licence version 3 (142). Open source licences dictate that when software incorporating open source components is distributed, a copy of the source code for the open source components should be distributed alongside the software or made

available in the same way. This does not apply to any data processed by the software. Notable examples of open source software include the R statistical package, the Mach Microkernel that forms the basis of MacOS, the Linux kernel and much of the rest of the operating system, and the PostgreSQL database package (upon which several medical record storage systems are built). There is no obligation whatsoever to share any data created, stored or analysed in any of these, or any other systems simply because they are completely or partly open source. There are consequently no data protection issues; the two concepts are unrelated.

4.4.5 Problems encountered when extracting data

4.4.5.1 Data format and layout issues

There were some issues encountered for iSOFT as well similar to iSOFT. For example there was an issue found where Read code descriptions and Read codes themselves were all in the same column rather than in separate columns. The problem was resolved by improving the regex notation used. Again continual checking and testing was the main method to identify and solve the data quality issues meant modifications to the form and dose regex definition lists and these are found in appendix 8.8. However once changes were made the issue was resolved and the missing values were correctly found in the extracted files.

4.4.5.2 Software programming issues

One of the key problems with iSOFT was that at the sole practice concerned there was only a PDF version of the complete EMRs available. Different approaches were taken and the problem is explained in detailed already in this chapter but the solution was to use a freely available software programme to act as the intermediary PDF-text converter. Another problem occurred when trying to build upon what was already done for systmOne as this was the second GP system to be catered for in the extraction software and was the module created for iSOFT EMRs. There were some initial problems as the software complier detected conflict errors when using systmOne. This was resolved by separating

the extraction software into modules catering for each type of GP system helped organise the software development. This greatly helped to ease the situation by separating the coding into these separate modules and allowing code checking to occur in a manageable and organised way.

4.4.5.3 The iSOFT Sections Identified Post Conversion

ISOFT records have a uniqueness that they can vary their structure according to the patient data that is to say that if for any reason there is no data present inside a subsection then that subsection does not exist. From a regex point of view it is a challenge as regexes thrive on consistency and structure. However after analysing iSOFT EMRs it was possible to identify headers that were always present these have been listed above; they also allow extraction of the key areas of demographics, medical and therapy files successfully. Post conversion the text file records were analysed and these were the four section headings identified in the converted EMRs from iSOFT:

- Patient – This contains the demographic information of the patient.
- All Medication – Contains all the drugs they have been prescribed.
- All Problems – All Read code diagnoses
- All Reminders – All reminders issued for the patient

The above four sections are extracted using the following regular expressions:

@"(<b>Single\sPatient\sSummary<\b>)";

@"(All\sMedication)";

@"(All\sProblems)";

@"(All\sReminders)";

Rationale:

- These subsection characteristics were consistently present within an iSOFT record whether or not other subsection headings were present. It therefore presents a stable view of the records in a situation where the structure of one iSOFT record can be slightly different to another EMR.

- Within these subsections the areas of covering a patient, medical and drugs can still be isolated. As a result it still allows benefits to be gained from increasing the accuracy of secondary regexes within each subsection when extracting specific data such as date of birth.

4.4.6 The iSOFT Demographics Extraction

The following demographic variables were identified from the demographics containing PatientRecord.txt subsection:

- Patient name
- NHS Id
- Date of birth
- Gender
- Address
- Marital status
- Registration Date
- Registered GP
- Postcode

4.4.6.1 Patient Name, Address and NhsId

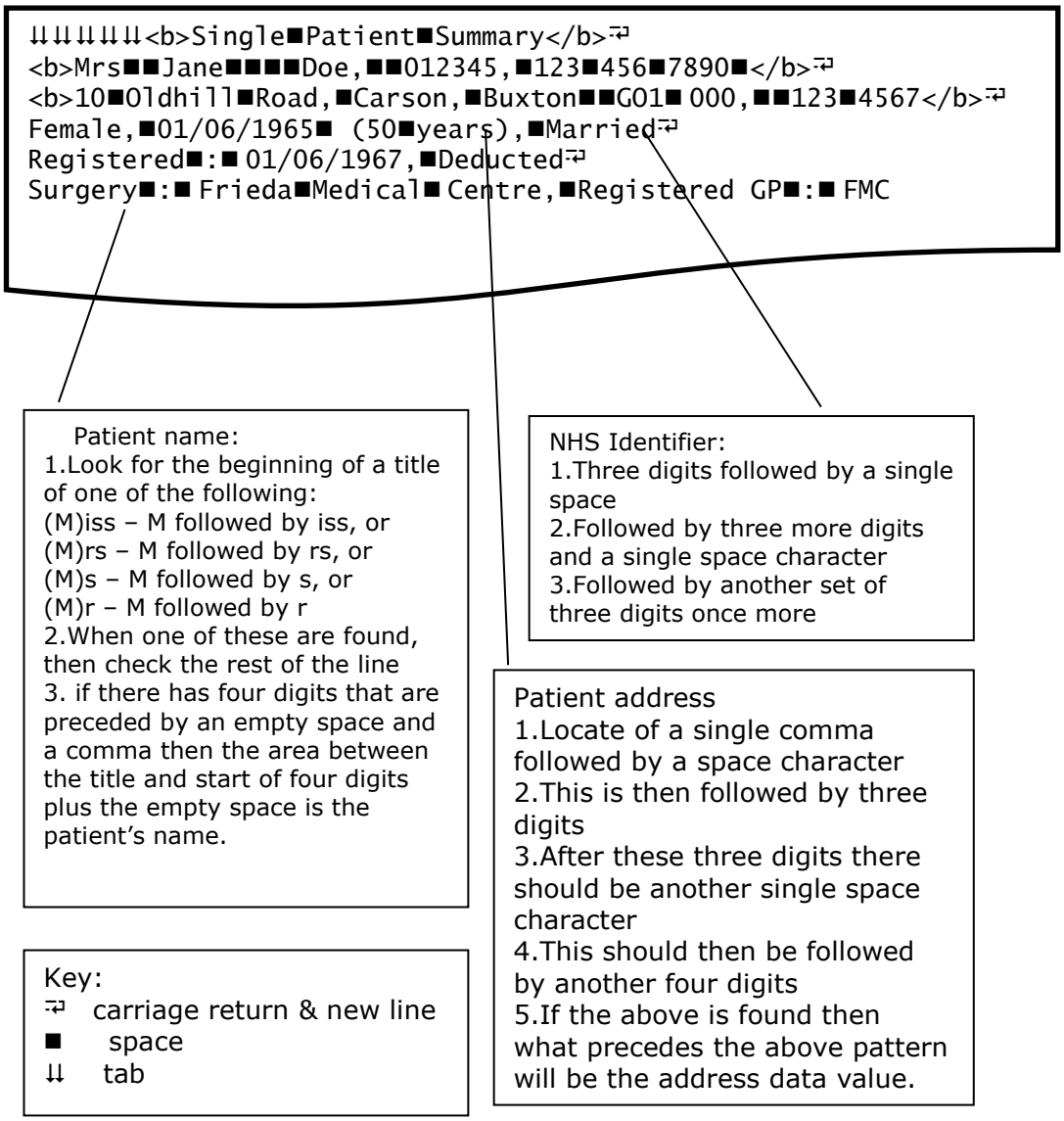


Figure 21 – iSOFT: extraction for Patient name, address and nhsId

4.4.6.2 The Patient Name

In iSOFT records there is no actual label that indicates the name of the person; instead the variable name in the context of its position relative to other items in patient information section can be used to find this data.

Patient name in iSOFT is extracted using the following regular expression:

```
@"(?<name>M[r|s|iss|rs].*(?=\,\s+\d{4}))";
```

Rationale:

- The name always starts after the heading 'Single Patient Summary' and is the first patient value we encounter.
- There is no clear label that marks it out where it finishes in relation to other commas found on the same line that indicate different data.
- Attempted to use an extraction that would have taken everything from very first character to very first comma and then produced a filter regex to remove the words 'Single Patient Summary' and any other non-sensible value. Did not use this approach as risked catching data that would need further filtering which may open extraction up to customised filtering depending on structure of records encountered.

4.4.6.3 The NHS Identifier (NHS ID)

This is unique for each patient and is allocated by the NHS itself and is extracted next. Due to its allocation by the health service, the NHS ID itself is consistently always made up of 10 digits for any patient.

NhsId is extracted using the following regular expression:

```
@"(?<nhsId>\d{3}\s\d{3}\s\d{4})";
```

Rationale

- The nhs identifier is not labelled as to that effect in the record the only way to detect its presence is the combination of spaces and length.
- There is also a local identifier next to the NHS Identifier so there was a need to be very specific about the location of spaces and number of digits.
- The length of the nhsID is not sufficient for iSOFT records as the GP software system software has punctuated it with spaces which the regex must be able to handle.

4.4.6.4 The Patient Address

The patient address is consistently placed near to specific sequence made of digits, spaces and commas. This address is made up of the full address including the postcode and always precedes a comma, three digits followed by a space and a further four digits which is the telephone number of the patient minus the area code.

In iSOFT address is extracted using the following regular expression:

```
@"^(?<address>(.)*)(?=\,\s+\d{3}\s\d{4})";
```

Rationale:

- The address is not marked in where it begins and where it finishes.
- Using the opening digits only is not sensible as a house number can be more than two digits or less or may not have any digits at all.
- It's important to take the whole line of address as it includes the postcode and this not separately placed or labelled.
- Using the digits to the right of the address is best way forward as these are consistently found in this location and can be used as a demarcation point for address end.

4.4.6.5 The Postcode

Postcode validation must adhere to the basic rules for UK postal codes. Essentially composed of two parts and a combination of five to seven letters and/or numbers.

In iSOFT the postcode is extracted using the following regular expression:

```
@"(Post\sCode([\s]+)\:\s)(?<postcode>.*)";
```

Rationale

- In this situation it was considered best practice to extract the postal code from the full address itself extracted rather than the record, this was due to the ambiguity that could possibly occur when the postcode is seen alongside the rest of the record.

4.4.6.6 Gender, Date of Birth, Registration Date

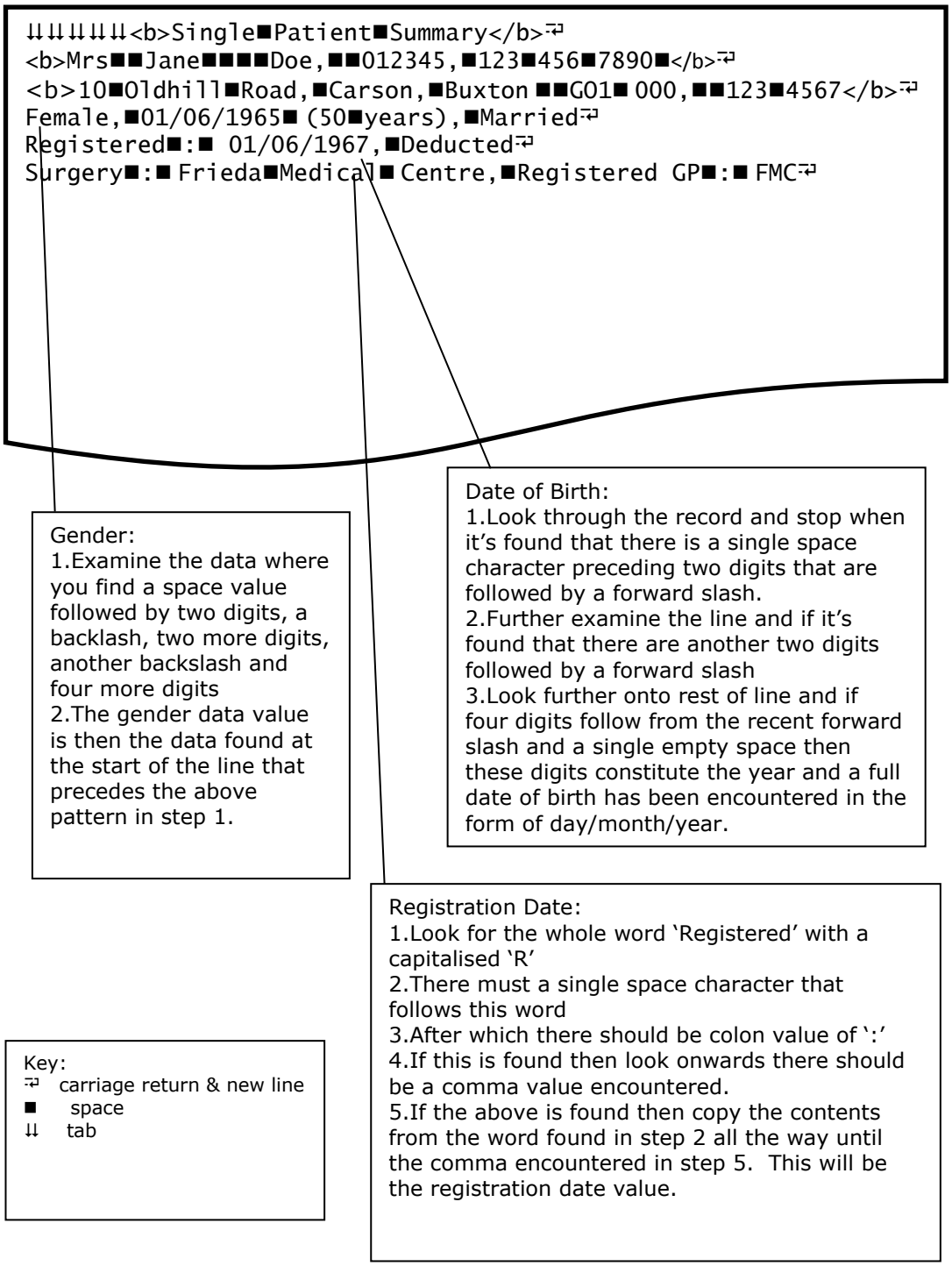


Figure 22 – iSOFT: extraction of gender, date of birth, registration and date

4.4.6.7 The Gender

The patient's gender information is always found near to a date value but not next to anything specific to indicate that the gender information follows on.

In iSOFT gender value is extracted using the following regular expression:

```
@"(?<gender>.*(?=\,\s\d{1,2}\/\d{1,2}\/\d{1,4}))";
```

Rationale

- The gender is the first value encountered after the NHS ID is followed by the date of birth, so is sandwiched between two values which extracted earlier.
- It shares a characteristic with the date of registration in that it begins a line and precedes a date value and is made of the alphabetic characters but is does not however contain a colon ':' preceding the date and so remains unique.

4.4.6.8 The Date of Birth (DOB)

The date of birth is structured in a unique way using a combination of digits, spaces and the forward slash character: '/'.

In iSOFT date of birth is extracted using the following regular expression:

```
@"(?<dob>\s\d{1,2}\/\d{1,2}\/\d{4}\s)";
```

Rationale

- This date is unique from other potential dates as they do not have the combination of single spaces in the same way. For example the registration date further along appears the same but there is the difference in spaces surrounding the dates.
- The DOB always occurs after the gender value and a comma so using this as its unique fingerprint it can be isolated from other dates which also

have the same general structure such as registration date which occurs on next line.

4.4.6.9 The Registration Date

This refers to the date that the individual registered at the practice as a patient. This is also expressed consistently in the same notation and is always found in the same place. This is important since there are other date values which share the same notation but refer to other bits of patient date such as date of birth and in some cases date of death.

In iSOFT registration date is extracted using the following regular expression:

@"(Registered\s\:\s)(?<registrationDate>(.)*)(?=\,)";

Rationale:

- The data value associated with the registration date does have its' own label; this is first demographic value encountered that has this feature.
- The most accurate way to isolate this value is to make use of its label notation within the regular expression itself.
- Using this approach also helps to differentiate it from other dates found with the same structure.

4.4.6.10The Registered GP and Postcode

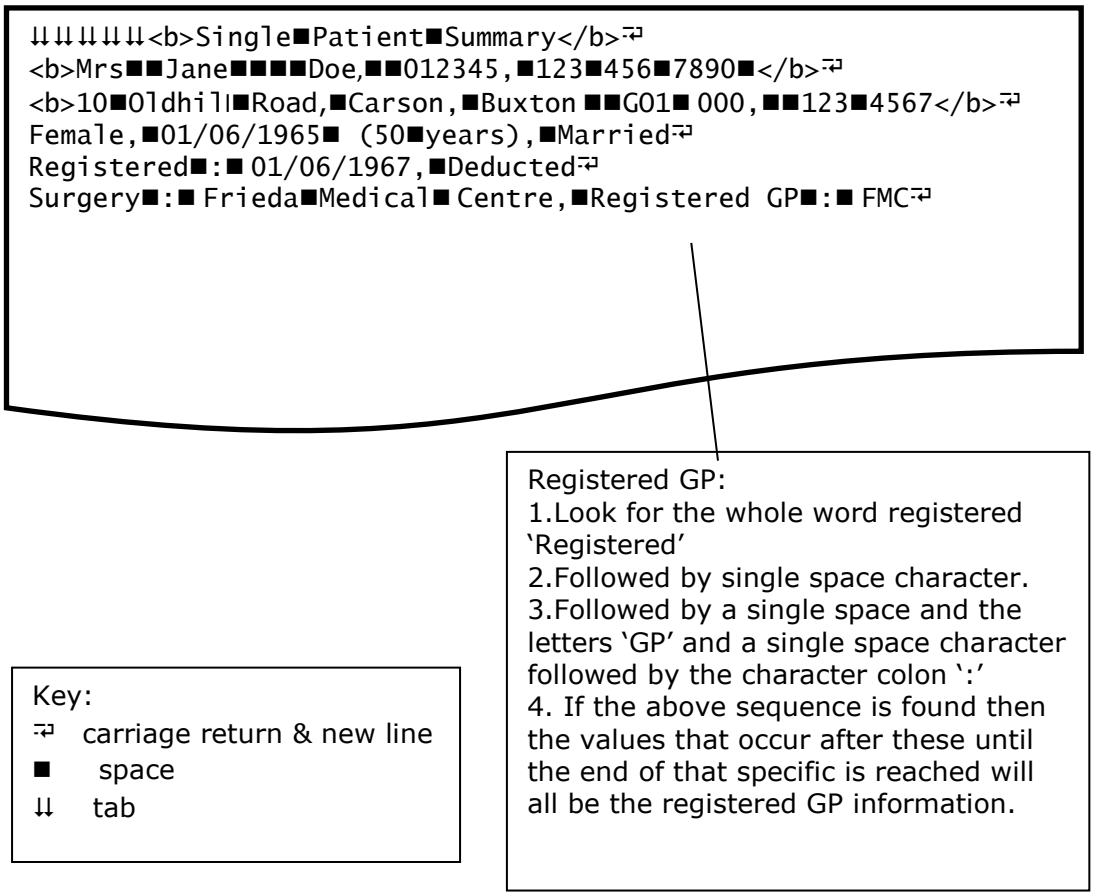


Figure 23 – iSOFT: extraction of registered GP and postcode

4.4.6.11The Registered GP

This refers to the general practitioner registered for the patient.

In iSOFT the registered GP is extracted using the following regular expression:

@"(?<registeredGp>(?<=\sRegistered\sGP\s\:).*)";

Rationale:

- This value also has its own label with the record structure. As a result it is the most accurate way of ensuring that what follows is the actually data value that we are interested.
- It was not sensible to use any titles such as 'Dr' in this regular expression as many practices do not indicate a specific individual as the GP but rather indicate an abbreviation of the practice name.

- Being at the end of line gives a useful and consistent end of data value marker rather than indicating an end of characters encountered.

4.4.7 The iSOFT Medical Extraction

The following section looks at extracting data for iSOFT's medical file. This aims to capture the date, Read code for the diagnoses, description of the Read codes and any other potentially useful information attached to each entry. Each entry in the medical refers to the recorded problem(s) that the patient presented to the GP with. In iSOFT records all such diagnoses are encoded in terms of Read codes.

4.4.8 Checking of Read or Clinical Codes

An extra step of validation of extracted Read codes was used in the same manner as described in systmOne. This was achieved by matching Read codes against a dictionary list of terms provided by the Health and Social Care Information Centre (HSCIC) so that the corresponding Read codes could be checked as being valid (see appendix 8.10 for more information).

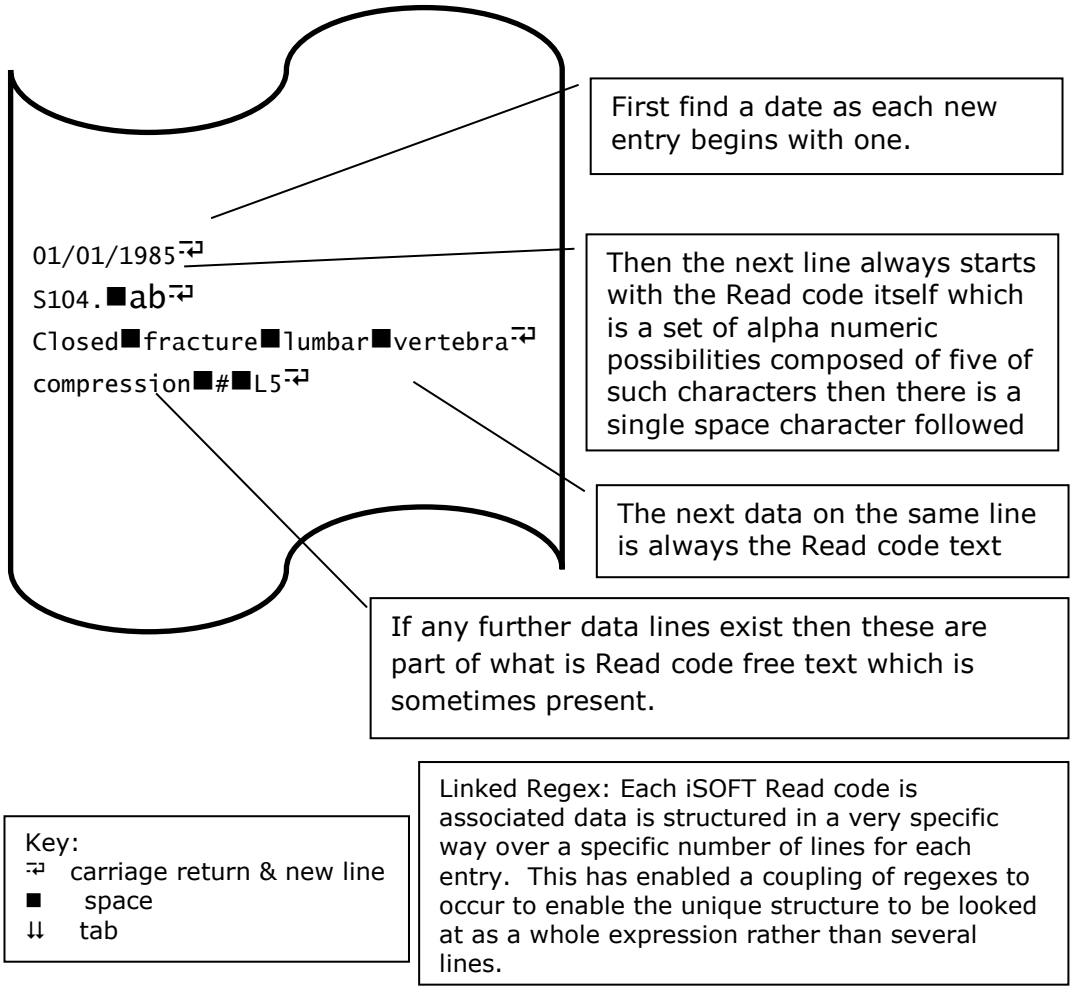


Figure 24 – iSOFT: extraction of medical date, Read code, Read code text and any free text

- Date – Each therapy received has a date associated along with it; so it becomes possible to define a pattern that matches the date the patient received the therapy indicated by the Read code.
- Read code – This is the Read code itself and represents it as a 5 digit unique identifier
- Read code Text – this is brief description of the Read code.
- Read code Free Text – This part refers to ad-hoc text inserted by the practice themselves or some other associated interesting comment by the doctor/practice staff.

In iSOFT records the medical data can be extracted using the following regular expression:

@"(?<date>^(\d{1,2}/\d{1,2}/\d{4}))(?<readcodeText>.*)$";

@"(?<readcode>^\@.*|^[A-Za-z0-9|\.]{5}\s)(?<staffName>(\w{2,3}|---))$";

Rationale:

- Each medical event has a clear structure that can be identified and extracted accurately.
- This consistent and accurate structure allows the ability to identify and capture free text

4.4.9 The iSOFT Drug Extraction

The next section looks at how drug-related information was extracted from iSOFT records. It revolves around finding a structure so that date the drug was prescribed, the full name, instruction, dosage and form can then be extracted.

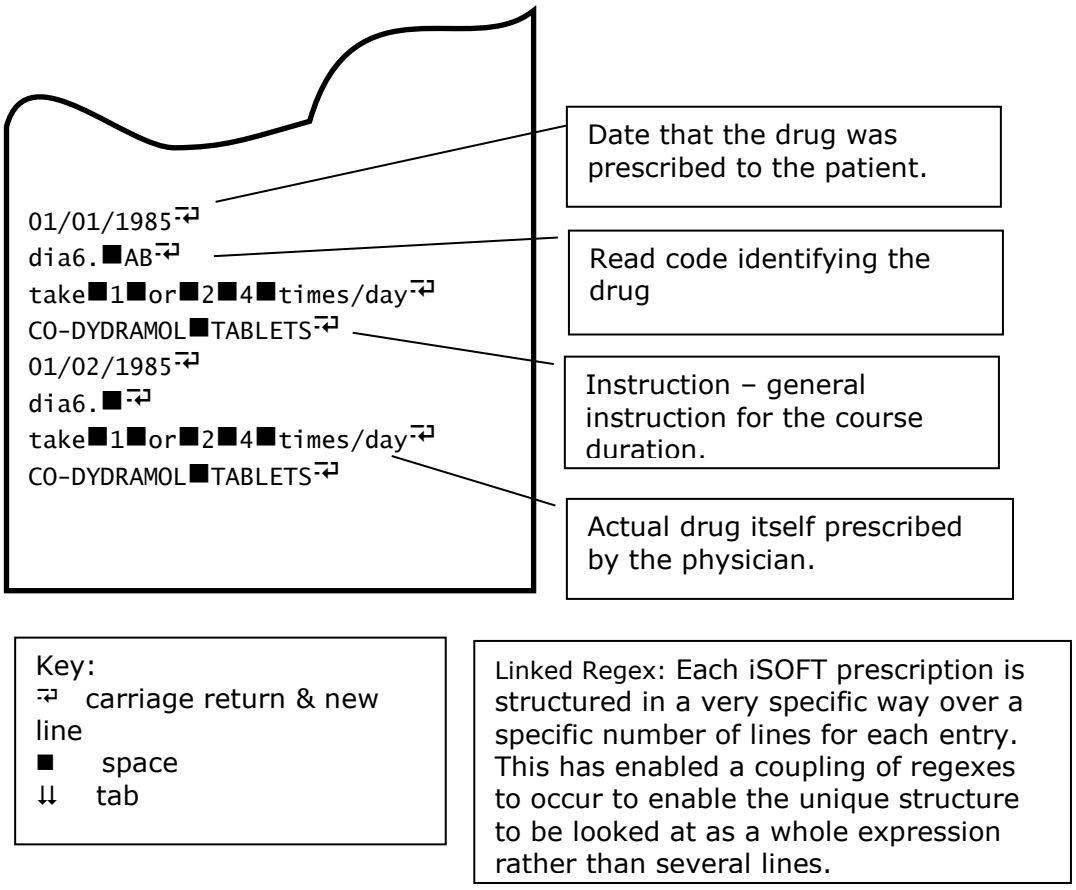


Figure 25 – iSOFT: extraction of prescription date, drug name, dose, form and instruction

Each iSOFT prescription is structured in a very specific way over a specific number of lines for each entry. This has enabled a coupling of regexes to occur to enable the unique structure to be looked at as a whole expression rather than several lines. Each set of lines is made up of the following components:

- Date of Drug Prescription – Date that the drug was prescribed to the patient.
- Drug Prescribed – Actual drug itself prescribed by the physician.

- Instruction – general instruction for the course duration.
- Read code – In ISOFT, Read codes are included as part of the prescriptions too.

In iSOFT drug data is extracted using the following regular expression:

@"(?<date>\d{2}/\d{2}/\d{4})\r\n

(?<readcode>.*?)\r\n(?<instruction>.*?)\r\n

(?<drug>(.|\r\n)*?)(?=\r\n\d{2}/\d{2}/\d{4})";

Rationale:

- All four key parts of: date, drug, instruction, read code occur in the same pattern for each entry presenting an opportunity to couple several regexes together in a single expression. It also ensures greater specific application of the regex to extract only the regex patterns contained within the single main expression.
- This allows a line by line extract for each entry in the journal and will automatically allow extractions to be relevant to entry itself.

4.5 Producing the Emis LV Data Files

4.5.1 Downloading the Emis LV files

All Emis LV records are received in a plain text file from practices using the Emis LV GP software system. See appendix 8.9 for download instructions.

4.5.2 Problems encountered when extracting data

4.5.2.1 Data format and layout issues

The first issues were Dates in Emis LV EMRs were again in a different format from previous GP systems encountered thus far. This is because although they were in the correct order and format in terms of digits they were separated by the dot operator rather than forward slashes and therefore the first step was to convert all the dates to the standard form. For example 1st January 1995 would appear as 01.01.2015 which was converted through the Extraction Software (ES) to 01/01/2015. Another issue encountered was that the order of a patient's full name, with surname occurring first and forenames follow on afterwards. This meant that when demographic files where put together the Emis LV patients names were not consistent with systmOne or iSOFT. A change was made in C# code to correct the issue by arranging Emis LV names to occur in the right order once initially extracted. The iterative and incremental process of evolving the regex trees so that forms and dosages were extracted continued for Emis LV. The regex form and dose lists are found in appendix 8.8.

4.5.2.2 Software Programming Issues

Another problem discovered was that Emis LV GP system downloaded EMRs did not contain Read codes (or any other clinical codes). It was expected from visiting previous practices and reading the history of Emis LV that Read codes or even locally created Read codes would exist in the EMRs. However only descriptions of Read codes were found in Emis LV

EMRs in the practices visited. However to enable effective analysis the corresponding code to the Read descriptions was still needed as well. This enables consistency when analysis is done on the rest of the EMRs that do contain Read codes like systmOne and iSOFT. Another reason is that during research it was found that whilst Read codes were seen in a consistent manner the corresponding descriptions were sometimes abbreviated at different practices. The solution adopted was to use the Read code description from EMRs where no Read code was present and the description was checked against a dictionary list of terms provided by the Health and Social Care Information Centre (HSCIC) so that the corresponding Read codes could be returned from the dictionary where the descriptions supplied from the EMRs matched. If a description failed then the code was instructed to move the description to free text. During testing the descriptions encountered matched the dictionary's terms so Read codes were successfully imputed and failures of Read term descriptions were all accounted for by moving to free text for manual review. At the time of writing the most recent, comprehensive and officially used clinical codes for the United Kingdom were downloaded using the Technology Reference data Update Distribution (TRUD) site which was owned and run by the Health and Social Care Information Centre (HSCIC) and since July 2016 this has been rebranded as NHS digital. The pack downloaded was used for searching for both Read version 2 and Clinical Terms Version 3 (CTV3) was the 15.0.0 release pack (141) (see appendix 8.10 for more information). Emis LV was the only GP software system that had the issue of Read descriptions without Read codes being present in the EMR. This has been discussed earlier in this chapter. Although the researcher was informed by staff at an Emis LV practice that the descriptions and codes are visible on the screen to GPs but during whole record rendering to move to another practice the Read descriptions are

present. A description of the dictionary from HGCIC can be found in appendix 8.10

4.5.3 Splitting the Emis LV file into subsections

The first to determine separate the record into sections. The top boundary refers to when the section starts and the bottom boundary refers to where that section completes. By assigning a numerical value to each of boundaries top and bottom it is possible to subtract them from one another leaving the remainder to be the subsection that we are seeking. This can be repeated for all subsections until the end of the record is encountered. An Emis LV record has the following subsections:

- Patient Summary – Demographic related data
- Active Problems - Recent diagnosis
- Allergies – Any allergies patient may have
- Diseases or Operations –
- Health Status - summary of health recordings
- Family History – outlining family diseases
- Present Medication – recent prescriptions
- Past Medication – past prescriptions
- Immunisations – any vaccinations are listed here
- Diary Entries Follow-up – listing any assessments that are due
- All Entries (Except) Values – Listing of all health problems for patient
- All Consultations – Catch all listing for anything not listed in 'All Entries'
- Biological Values – Other health measurements not covered in health status section.

Having identified the structure that makes up an Emis LV record it was then necessary to express the structure as a series of initial regexes that 'cut up' the record into separate files for secondary regexes to work on. The following rules apply when identifying the subsections marked above:

- Look for the word 'PATIENT' followed by a single space and the word 'SUMMARY'. Assign a unique numerical value for this point once it has been reached and remember this value (rules continue below).
- Look for the word 'ACTIVE' followed by two single spaces and the word 'PROBLEMS'. Assign a unique numerical value to this point once it has been reached and remember this value.
- Look for the word 'ALLERGIES ' and assign and unique numerical value once it has been encountered and remember this value.
- Look for the word 'DISEASES' followed by a single space the word 'OR' another single space and the word 'OPERATIONS'. Assign a unique numerical value to this point once it has been encountered and remember this value.
- Look for the word 'HEALTH' followed by a single space and the word 'STATUS' and another single space. Assign a unique numerical value to this point once it has been encountered and remember this value.
- Look for the word 'FAMILY' followed by a single space and the word 'HISTORY'. Assign a unique numerical value to this point once it has been encountered and remember this value.
- Look for the word 'PRESENT' followed by a single space character and the word 'MEDICATION'. Assign a unique numerical value to this point once it has been encountered and remember this value.

- Look for the word ‘Past’ followed by a single space and the word ‘Medication’. Assign a unique numerical value to this point once it has been encountered and remember this value.
- Look for the word ‘IMMUNISATIONS’ and assign a unique numerical value to this point once it has been encountered and remember this value.
- Look for the word ‘DIARY’ followed by a single space, the word ‘ENTRIES’ followed by a single space, the word ‘AND’ followed by a single space, then the word ‘FOLLOW’ and a single space and finally the word ’UP’ followed by a single dot character. Once this has been encountered assign a unique numerical value to it and remember this value.
- Look for the word ‘ALL’ followed by a single space, the word ‘ENTRIES’ followed by a single space, and opening bracket then word d ‘EXCEPT’ followed by a single space, the word ‘VALUES’ followed by a closing bracket. Assign a unique numerical value to it and remember this value.
- Look for the word ‘All’ followed by a single space and the word ‘consultations’. Assign a unique numerical value to this and remember this value.
- Look for the word ‘Biological’ followed by a single space and the word ‘Values’. Assign a unique numerical value to this and remember this value.

Based on the above rules an Emis LV record is therefore extracted using the following regexes:

```
@"(PATIENT\sSUMMARY)"
@"(ACTIVE\s\sPROBLEMS)";
@"(ALLERGIES)";
```

@"(DISEASES\sOR\sOPERATIONS)";

@"(HEALTH\sSTATUS\s)";

@"(FAMILY\sHISTORY)";

@"(PRESENT\sMEDICATION)";

@"(Past\sMedication)";

@"(IMMUNISATIONS)";

@"(DIARY\sENTRIES\sAND\sFOLLOW\sUP\.)";

@"(ALL\sENTRIES\s\(EXCEPT\sVALUES\))";

@"(All\sconsultations)";

@"(Biological\sValues)";

Rationale

- Once numerical values have been assigned to the headings identified in an Emis LV record and numerical values have been assigned to the start of the patterns above then it becomes possible to use these numerical values to find the numerical difference from one another to ensure a consistent an accurate subsection extract for each regex above .

- It is better to use a numerical value to identify start of the subsections so that there is no possibility that there could another set of words that could be confused later on in the whole record. For example there won't be the words 'Medication' causing a conflict between present and past medication as a numerical value is indicated for this regex and it is this value rather than the words that are used in the subtraction that causes the splitting off of these subsections.

4.6.1 The Emis LV Demographic Extraction

The following section outlines how demographic related information that forms the Patient file is achieved. Demographic related data can items

such as name, nhsID, address and date of birth amongst other potentially useful items.

4.6.1.1 EMIS LV Demographics: Name, Address, Date of Birth

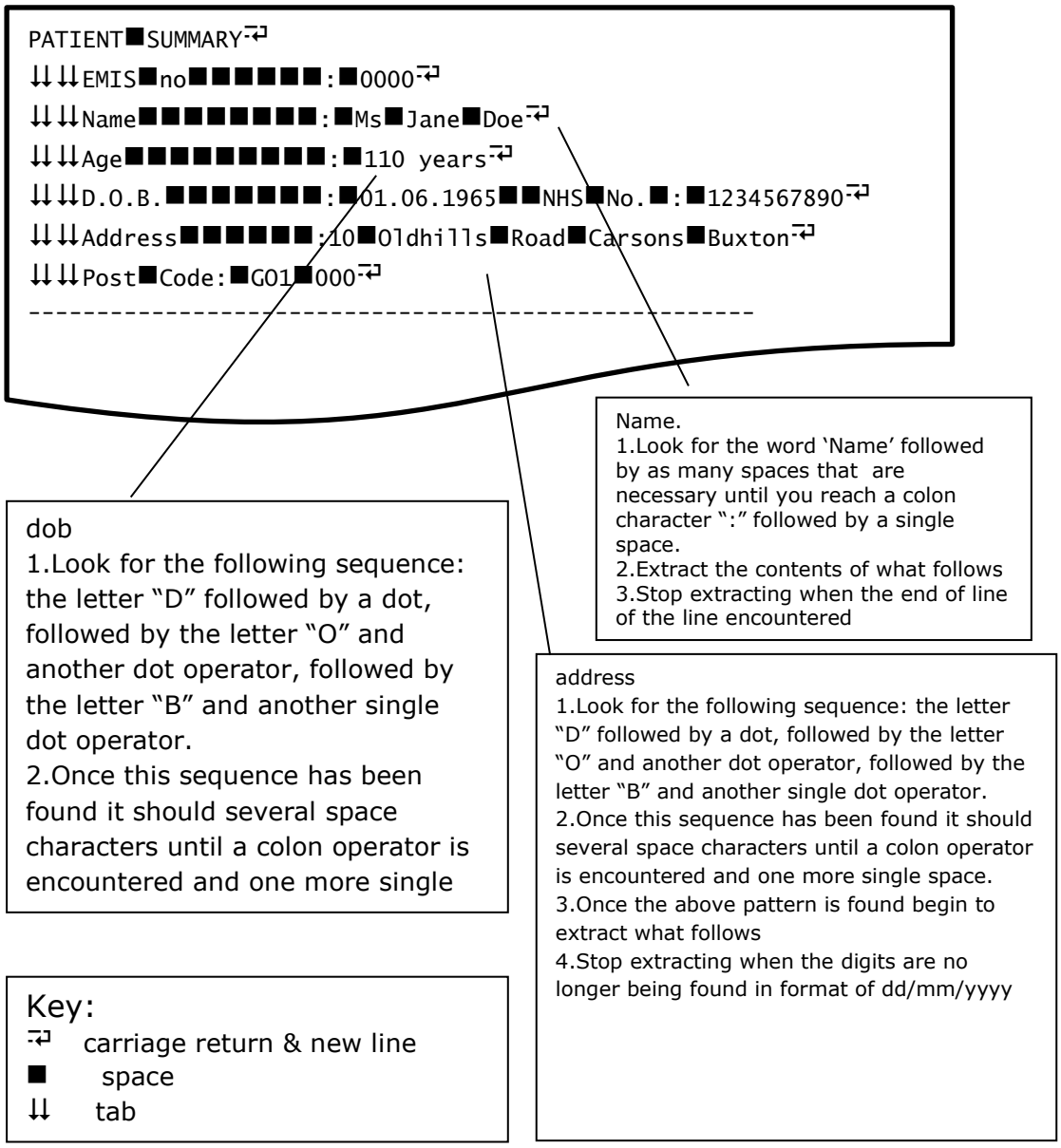


Figure 26 – Emis LV: Extraction of name, address & date of birth

4.6.1.1.1 The Patient Name

This refers to the name of the patient. In Emis LV records name is extracted using the following regular expression:

@"(?<name>(?<=Name\s+\:)\s.*)";

Rationale

- The variable of name has its own label associated with it however it's quite far from the data itself separated by spaces and a colon character.
- This approach allows the data to be extracted rather than the label as it uses a positional match using the label itself ensuring accuracy and consistency.

4.6.1.1.2 Date of Birth

This variable refers to the date of birth of the patient.

In Emis LV records DOB is extracted using the following regular expression:

@"(?<dob>(?<=D\.O\.B\.\s+\:)\s\d{1,2}\.\d{1,2}\.\d{4})";

Rationale

- The date of birth is preceded by its own unique label but is separated by one or more space characters. Using this unique positional feature we can extract what is found to the right of the sequence so that the actual date sequence is encountered and not the label itself.
- This is a unique line as it contains both the date of birth and NHS number. This means that it is not possible to continue extracting

the rest of the line after date of birth label and it shares spaces with the NHS number nearby.

- As a result it is best to specify in detail the number of digits so that extraction knows that it will cease its operation before the NHS number.

4.6.1.1.3 Address

This information refers to the permanent residence of the patient.

Regex to implement the above

- @"((Address([\s]+)\:\s)(?<address>.*))";

Rationale

- This data item has its own unique label associated with it though separated by a quite a few empty spaces it is possible to express that it takes as many as necessary which is the best way rather than link with a specific number of spaces that could change across records.

- The entire line that follows the pattern is populated by the address value and nothing else which means it’s better to take the rest of line until the end of line is found as addresses can have varying lengths depending on how the data has been input which is more likely not done in a uniform way across practices.

4.6.1.2 EMIS LV Demographics: Emis Number and NhsID

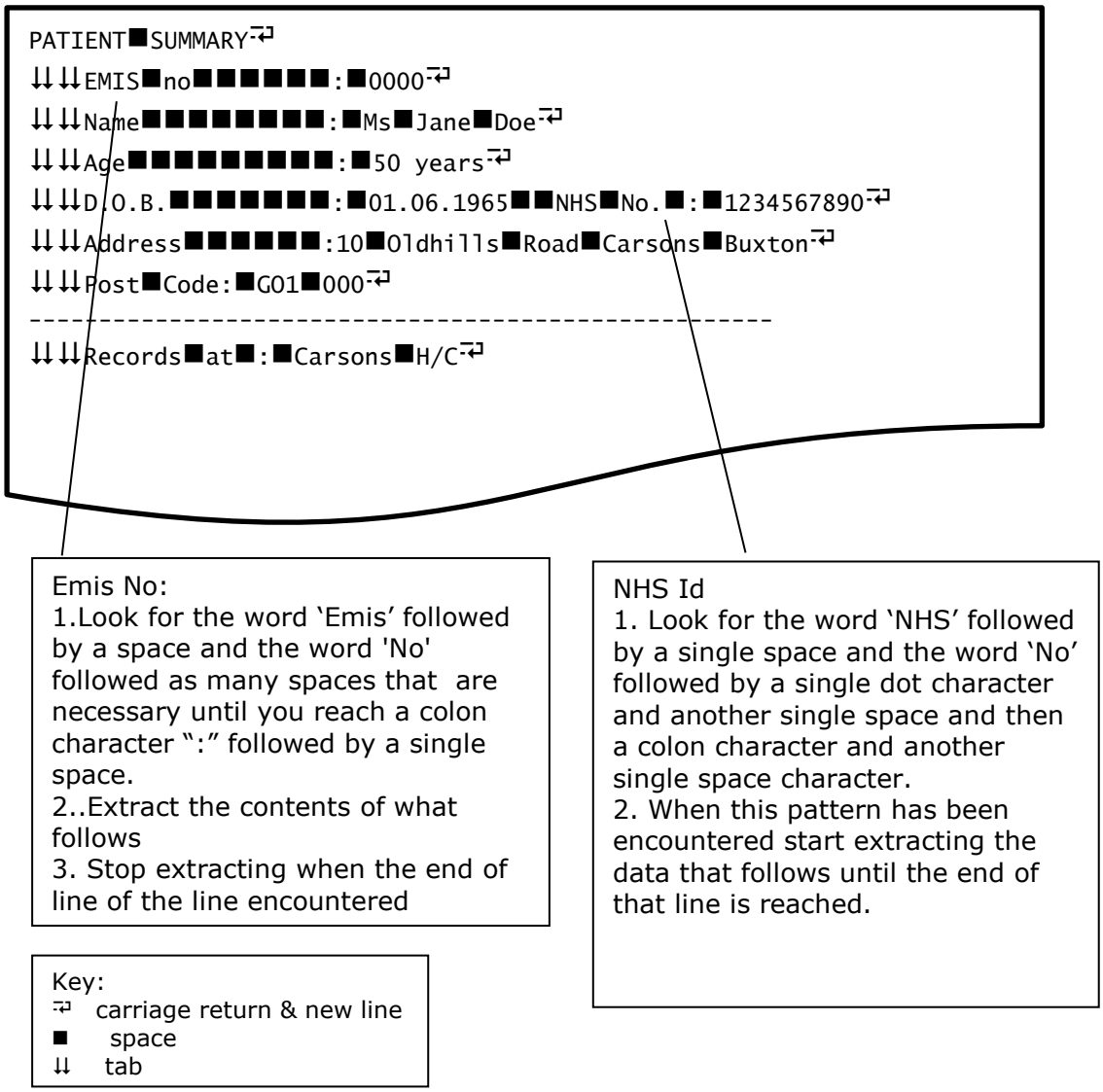


Figure 27 – Emis LV: Extraction of number and NHS Id

4.6.1.2.1 NHS ID

In Emis LV records NHS ID is extracted using the following regular expression:

@"(?<nhsId>(?<=NHS\sNo\.\s\:\s)\d{10})";

Rationale

- The NHS number has its own unique label structure and this label can be used as a positional marker so that the data to the right of it can be extracted successfully.
- In previous approaches it is possible to indicate a maximum number of digits to extract however in this case the data that follows the above data is the complete number sequence and then the end of that line is encountered.

4.6.1.3 Emis Number

This is specific to Emis GP systems and is the unique identifier within this system

Rationale:

- The Emis id is comprised of 4 digits however it is still better to use the label that precedes it in case Emis decide to change the length of their identifier.

Emis No is extracted using the following regular expression:

```
@"(?<emisNo>(?<=EMIS\sno\s+\:).*)";
```

4.6.1.4 EMIS LV Demographics: Postcode and Records at

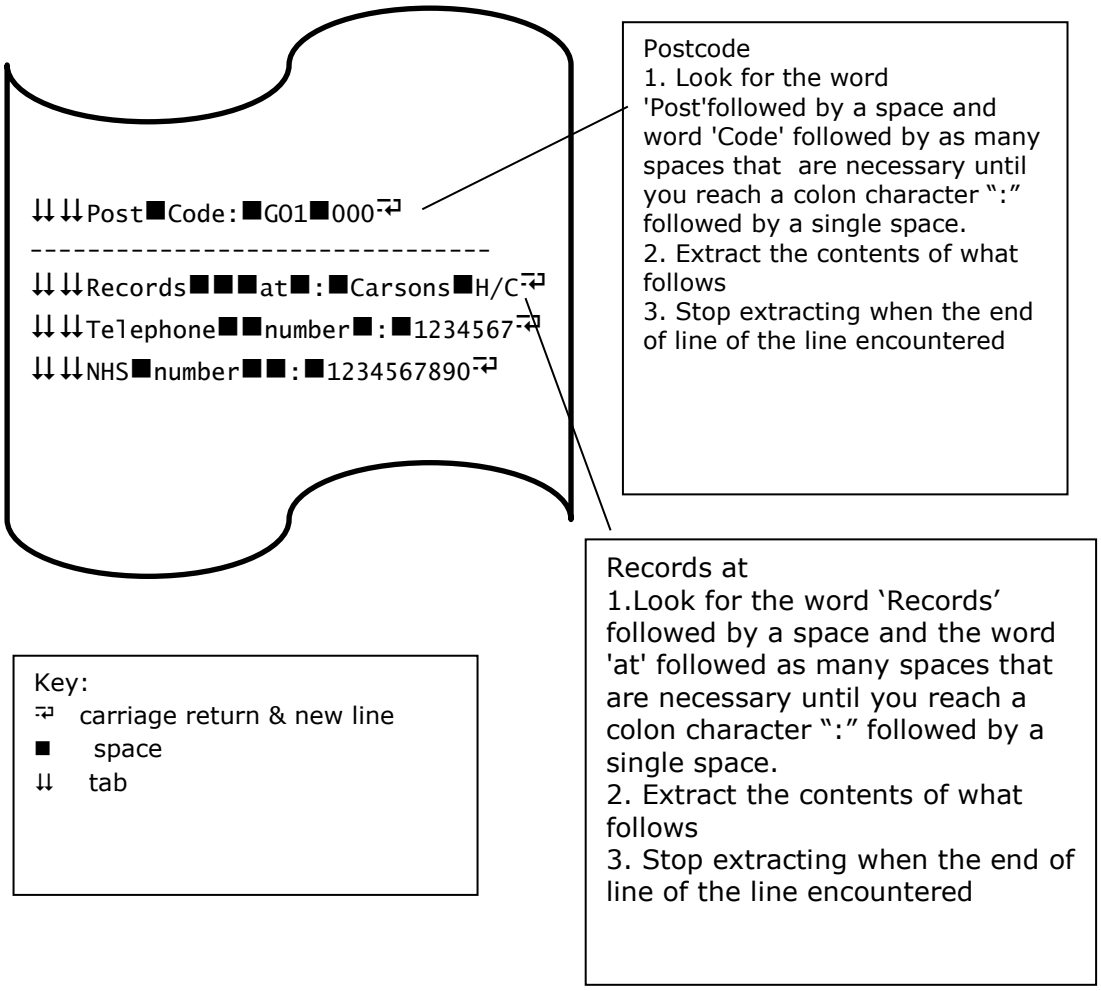


Figure 28 – Emis LV: extraction of postcode and records at

4.6.1.4.1 Postcode

This is the patient's postcode of their address.

Regex to implement the above

@"(Post\sCode([\s]+)\:\s)(?<postcode>.*)";

Rationale

- Postcode is not supplied as part of the address field of the patient but is given underneath it in the actual record. There was an option to take line beneath as well and include it in the postcode but in the interests of consistency and any other GP software system systems it's best to extract this same way using its label identifier and the sequence given above.

4.3.4.6 Records At

This refers to where the GP records of the Patient are kept. It also refers to where the paper records are stored.

Rationale:

- In some cases a practices can be part of a larger group, some in different places of the same town and paper records may be stored at a larger one for example. As a result greater accuracy lies in using the label rather than writing code that can match the practice name to the one being sought in the record. It is possible to do it in that manner but if there are minor variations in names being matched a false error may occur.

Records At is extracted using the following regular expression:

@"(?<recordsAt>(?<=Records\sat\s+\:).*)";

4.6.1.5 Hospital Number, Trading and Registering Dr

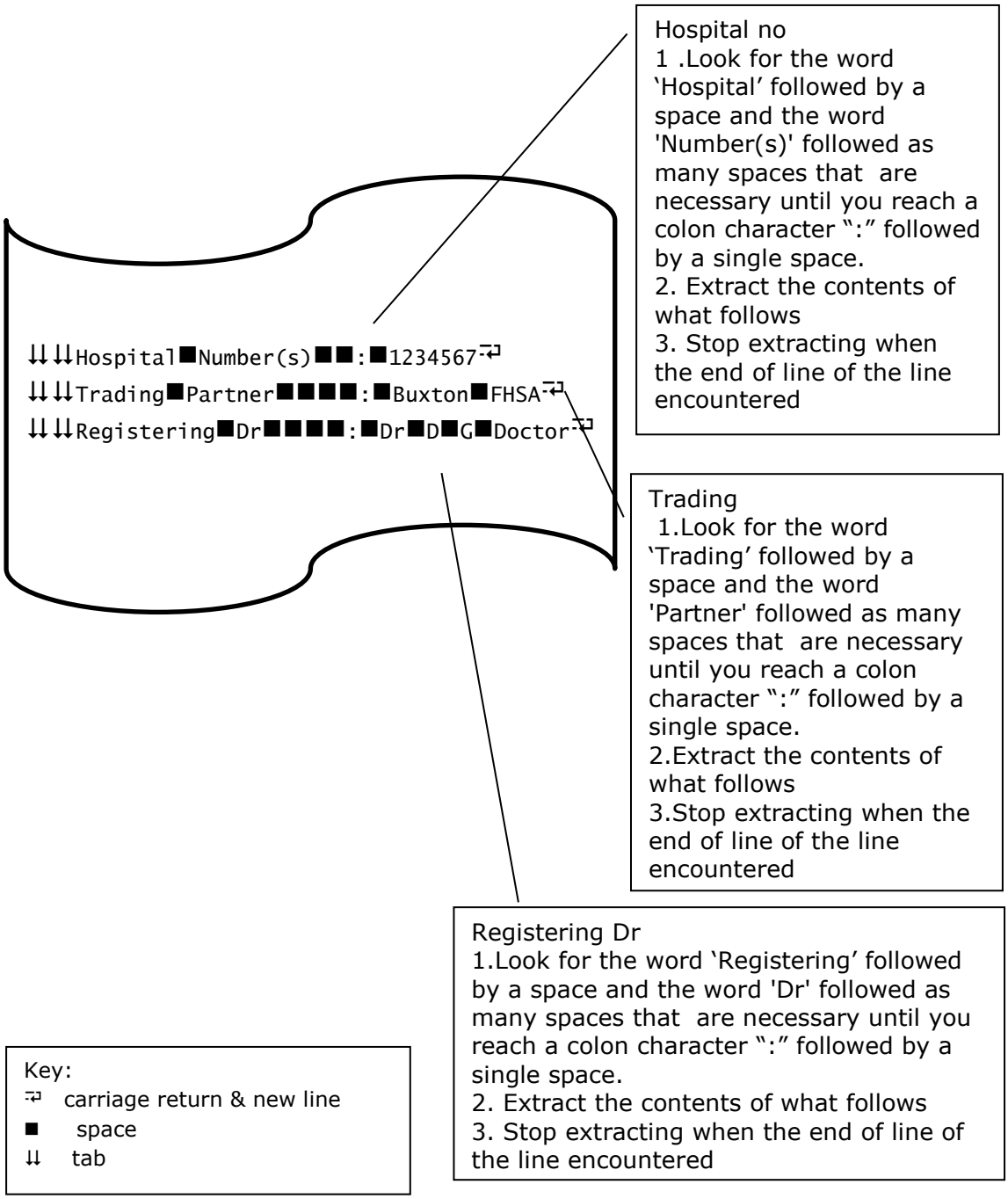


Figure 29 – Emis LV: Extraction of hospital number, trading and registering Dr

4.6.1.5.1 Hospital Number

This refers to the hospital number of the patient. The nhsID is still the universal identifier for health care in the UK but if a hospital number is present it will be extracted.

Rationale

Although the length and formatting of the hospital number means a regex can be designed if these things are modified they could conflict with the NhsId. To avoid this occurring the label that precedes this variable will do the work of highlighting the position to the regex.

Hospital number is extracted using the following regular expression

```
@"(?<hospitalNo>(?<=Hospital\sNumber\(s\)\s+\:).*)";
```

4.6.1.5.2 Trading Partner

Some GP practices list the trading partner in the records and this refers to the practice's financial relationship with the UK health service in the local environment. For example NHS commissioning board.

Rationale

- Trading relationships can change and this would need updating in the records. One option implemented was to analyse all the Emis records and use coding to match whichever was in the list of trading partners. However if a practice's trading relationship changes it's possible that a custom list may not include the name. It made much more sense to use the label as that is least likely to be modified compared to the actual trading partner.

Trading is extracted using the following regular expression

```
@"(?<tradingPartner>(?<=Trading\sPartner\s+\:).*)";
```

4.6.1.5.3 Registered GP

This is the registered GP that the patient is registered with.

@"(Registering\sDr([\s]+)\:\s)(?<registeredGp>.*)";

Rationale

- This data item has its own label associated with it and is conforms to the normal pattern of the label itself.
- It also has the trait where it is the only data item on that line and does not share space with other data items so it is a safe approach to extract the contents until the end of the line is reached. This is more effective than for example copying the next two words after 'Dr' with the aim being the Dr's first name and surname as this field may express data such as partners as it may or may not refer to the actual GP providing care.

4.6.1.6 Usual Dr, Dispensing and RPP road miles

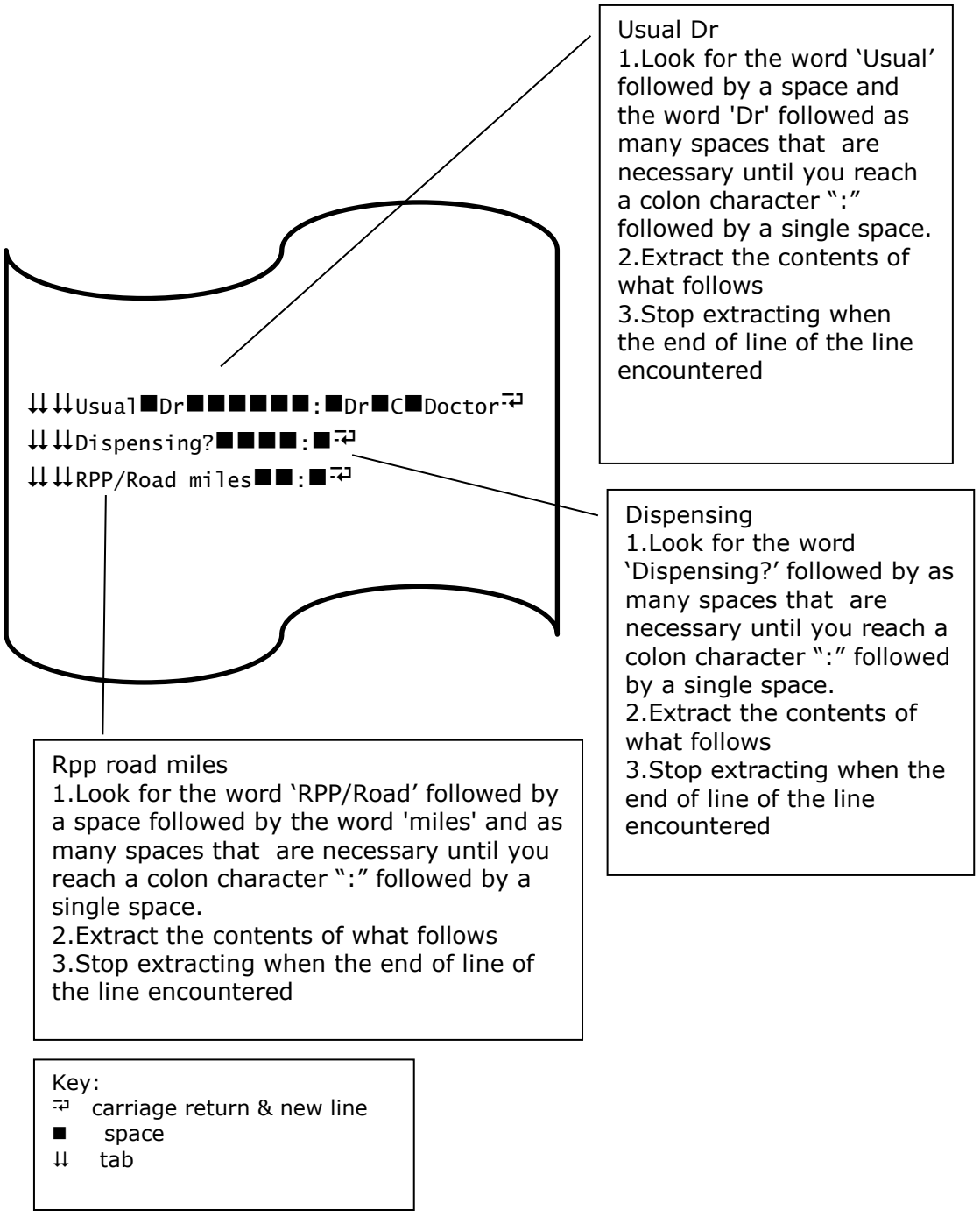


Figure 30 – Emis LV: Extraction of usual dr, dispensing & RPP road miles

4.6.1.6.1 Usual GP

This refers to the usual GP that the patient is registered to see.

Regex to implement the above

@"(Usual\sDr([\s]+)\:\s)(?<usualGp>.*)";

Rationale

- The word ‘Dr’ occurs several times in the same section of patient demographics to avoid conflict it is more robust to ensure the word ‘Usual’ is used as it adds uniqueness and also forms part of the label that identifies the specific data item.
- The line that contains this data does not share its space with any other data item so it is safe to extract the remainder of the line until the end of that is encountered.

4.6.1.6.2 Dispensing

This refers to whether the practice is also able to dispense the medicine it prescribes.

Rationale

Some practices may introduce this service or discontinue and although the response to this can be range of options to yes or no or left blank as has been seen in the majority of records the safest and long lasting approach is to use the identifier label that precedes where the value would ordinarily reside.

Dispensing is extracted using the following regular expression:

@"(?<dispensing>(?<=Dispensing\?\s+\:).*)";

RPP Road miles

This refers to the distance calculations associated with a patient. For all Emis LV records encountered so far in the study this has remained empty. However in case it is ever populated with potentially useful data a regex is still designed to handle it.

Rationale

- Since all Emis LV records encountered so far have this as an empty value it is not possible to design a regex based on formatting or appearance.
- It is also unknown that whenever it is present whether it would conflict in any way with an existing regex. As a result it's better to use the label which is always present irrespective of the associated value.

RPP Road miles is extracted using the following regular expression:

```
@"(?<rppRoadMiles>(?<=RPP\/Road\smiles\s+\:).*)";
```

4.6.1.7 Walking Quarters, Residential Inst. and Special Blocked

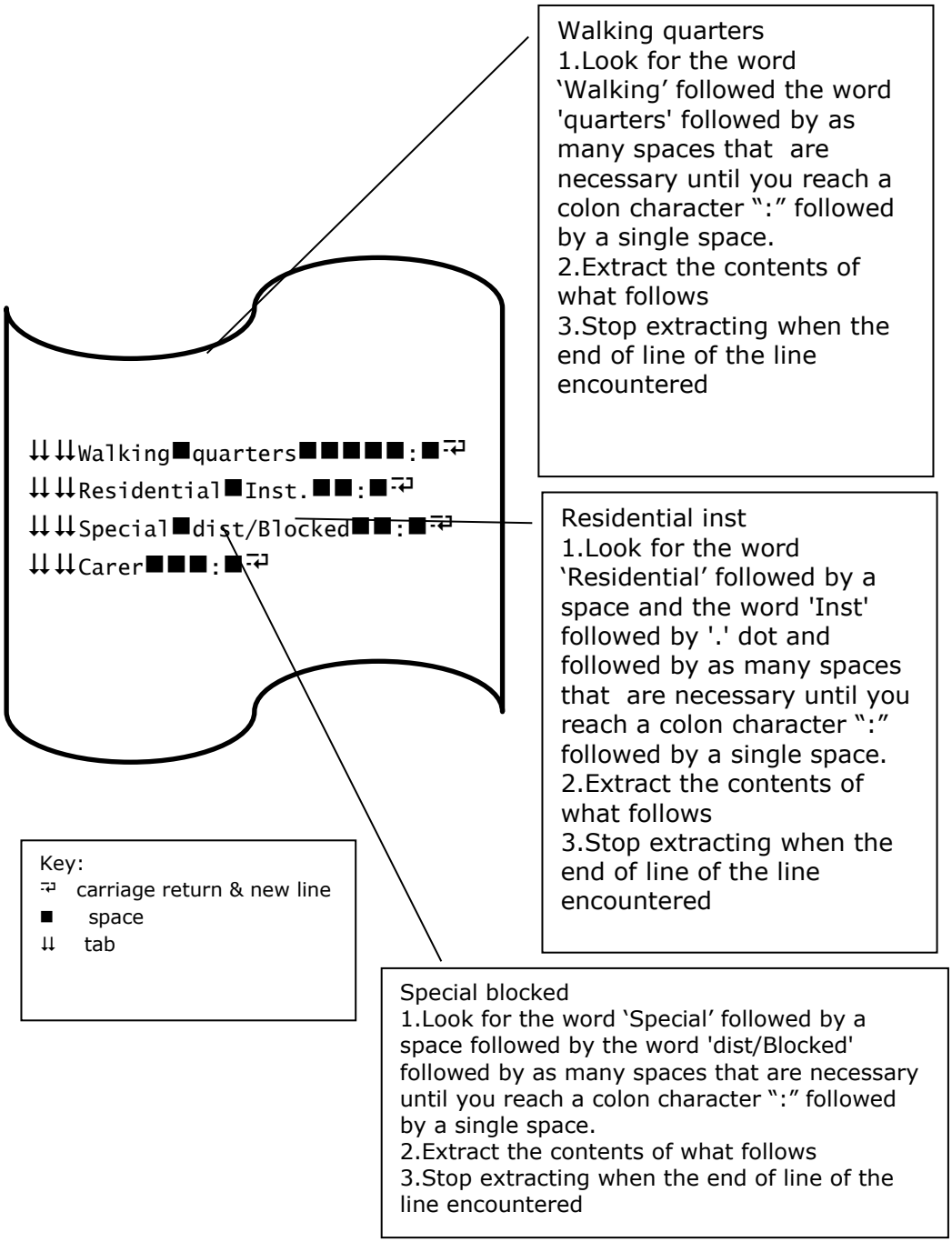


Figure 31 – Emis LV: Extraction of walking quarters, residential inst. & special blocked

4.6.1.7.1 Walking Quarters

This refers to a walking quarters measurement associated with patient. It is a measurement label found in all Emis LV records encountered and has been found to be empty thus far. It will however have a regex to handle any items if they are ever present. Walking quarters like RPP/Road miles gives an idea of distance measurements but since it is has always been empty thus far it cannot be said for sure what metric it is applied with.

Rationale

- Even though it has been seen to be empty in case Emis LV records not knowing anything about what possible metrics or formatting could be used it is sensible to utilise a regex that can handle the scenarios. It is better for long term functionality that a regex attempt to extract any data values associated.

Walking quarters is extracted using the following regular expression:

```
@"(?<walkingQuarters>(?<=Walking\squarters\s+\:).*)";
```

4.6.1.7.2 Residential Inst.

This refers to the patient's instruction regarding their residential data. Again like the RPP Road miles, Walking quarters the Residential Inst. has been found to be empty in all Emis LV records so far. It makes it difficult to ascertain the use of these fields in Emis LV records but nonetheless it's important for accuracy purpose that an attempt to extract should it exist.

Rationale

- Without observing real-life examples the specifics of Residential Inst. are unknown as to what it is composed of and looks like. However it maybe that patients may indicate some instructions in

this field at some point. Therefore makes sense to build that into the extraction process in case.

Residential Inst. is extracted using the following regular expression:

@"(?<residentialInstr>(?<=Residential\sInst\.\s+\:).*)"

4.6.1.7.3 Special dist/blocked

This refers to the data item associated with the Special dist/blocked value. Again it is not clear as to what this maybe constituted as due to the value always being empty thus far in all Emis LV records encountered. However to be sure of any changes in the long term designing a robust solution in case is a much better than to leave it altogether assuming it always be empty.

Rationale

- Structuring the regex to hitherto unknown value for Special dist/blocked would not be robust design. It is still vital to attempt to extract even though it has never been encountered in any Emis LV record thus far. By using the label structure it will ensure this happens without any unforeseen issues.

Special dist/blocked is extracted using the following regular expression:

@"(?<Specialdist>(?<=Special\sdist\/Blocked\s+\:).*)";

4.6.1.7.4 Registration Date

This is the date that the patient registered as a patient at this practice.

Regex to implement the above

@"(?<registrationDate>\d{1,2}\.\d{1,2}\.\d{4}(?=\sRegistration\sstatus\s))";

Rationale

- In this instance it is the only time that label associated with its data item is found to the right of the data itself.
- Previously in the above examples the structure has been that the label is followed by data not that data is followed by its label. For this reason it is better that the label is used as positional marker but this time the data to the right is extracted.
- It is better design to use the specific format of the date and dots shown as this line is preceded by an empty space which does not happen on other lines within the demographic section. The concern here is that it may be there in case any of the lines above overrun. As a result specifically expressing a date notation will ensure robustness in the face of any lines which may overflow above.

4.6.1.7.5 Registration status

This refers to the registration status of the patient. It's made of alphabetic characters and is always present in the records. It can have different values depending on the patient's circumstances when registering at the practice.

Rationale

- Registration status has been usually present in the Emis LV records encountered thus far. Being composed of alphabetic words and spaces it's far more effective to design for a mixture of words and digits and/or other characters as these may be present at some point. As a result the label that precedes the registration status will be used to identify the data item rather than from a custom range of the values that have been found.

Registration status is extracted using the following regular expression:

@"(?<regStatus>(?<=Registration\sstatus\s+\:).*)";

4.6.1.7.6 Carer

This refers to whether the patient has a carer recorded in their records. It is found in some records and composed of alphabetic characters and spaces.

Rationale

Based on findings so far the carer is the full name of the individual who is recorded as being the carer. As a result names can vary and/or be hyphenated or have nick names using inverted commas. Given the potential variation in values the carer label itself will designate the position of the actual carer data value. This ensures an efficient translation of the full name without missing any parts due to variation that might be encountered along the way.

Carer is extracted using the following regular expression:

@"(?<carer>(?<=Carer\s+\:).*)";

4.6.2 The Emis LV Medical Extraction

The following section described how Read code-related information was extracted from Emis LV records. This covered the areas of the actual Read codes for diagnosis, name of Read code, the date and any free text.

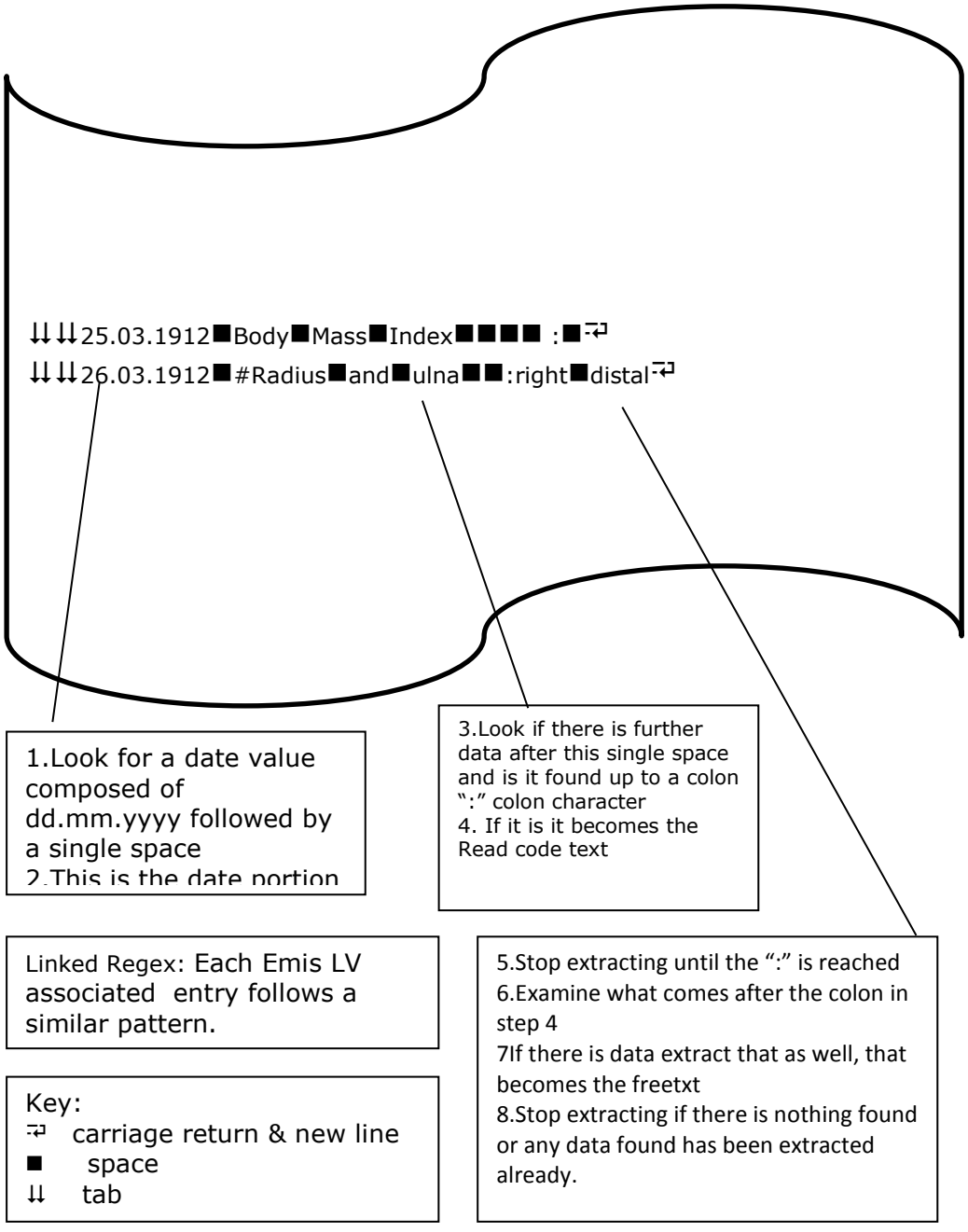


Figure 32 – Emis LV: extraction of medical information of date, Read code text and any free text

This refers to extracting the Read codes its associated data to examine the nature of the therapy provided to the patients. In Emis LV records the

entries follow a specific grouped pattern to indicate each new entry. There are however no Read codes present in Emis LV records only the Read code descriptions themselves.

4.6.2.1 Regexes to implement the above

```
@"(?<date>((\d{2}\.)?(\d{2}\.)?\d{4}))\s(?<readcodeText>(.*?)):(?<readcodeFreeText>.*)";
```

Rationale

- The entries are structured such that they always begin with a date. This allows isolation of a potential line match.
- By further analysing what comes after date we can be sure that a line containing Read code description and attached data has been encountered. It avoids the need to explicitly look for a Read code because the structure of Read codes can be ambiguous and they also absent from Emis LV records.
- By isolating the Read code description we can use this data item to extract a Read code by matching it to an external dictionary of terms.

4.6.2.2 Cleansing the Read codes

In order to gain Read codes from an Emis LV record it is necessary to match the extracted Read code descriptions to an external mapping dictionary from TRUD introduced earlier in this chapter. This allows any matching Read code descriptions to then pull out the corresponding Read code matches into the same extracted file thereby completing the picture with matched Read codes. It is necessary to use their definitions list as this group represents the UK authority on UK clinical codes. The extracted file is matched against this dictionary list of terms and matches returned in

the form of Read codes inserted into the file. The following section details how secondary regexes were applied to extract prescription-related information from Emis LV records. This covered the prescription's name, dose form and instructions to patients.

4.6.3 The Emis LV Drug Extraction

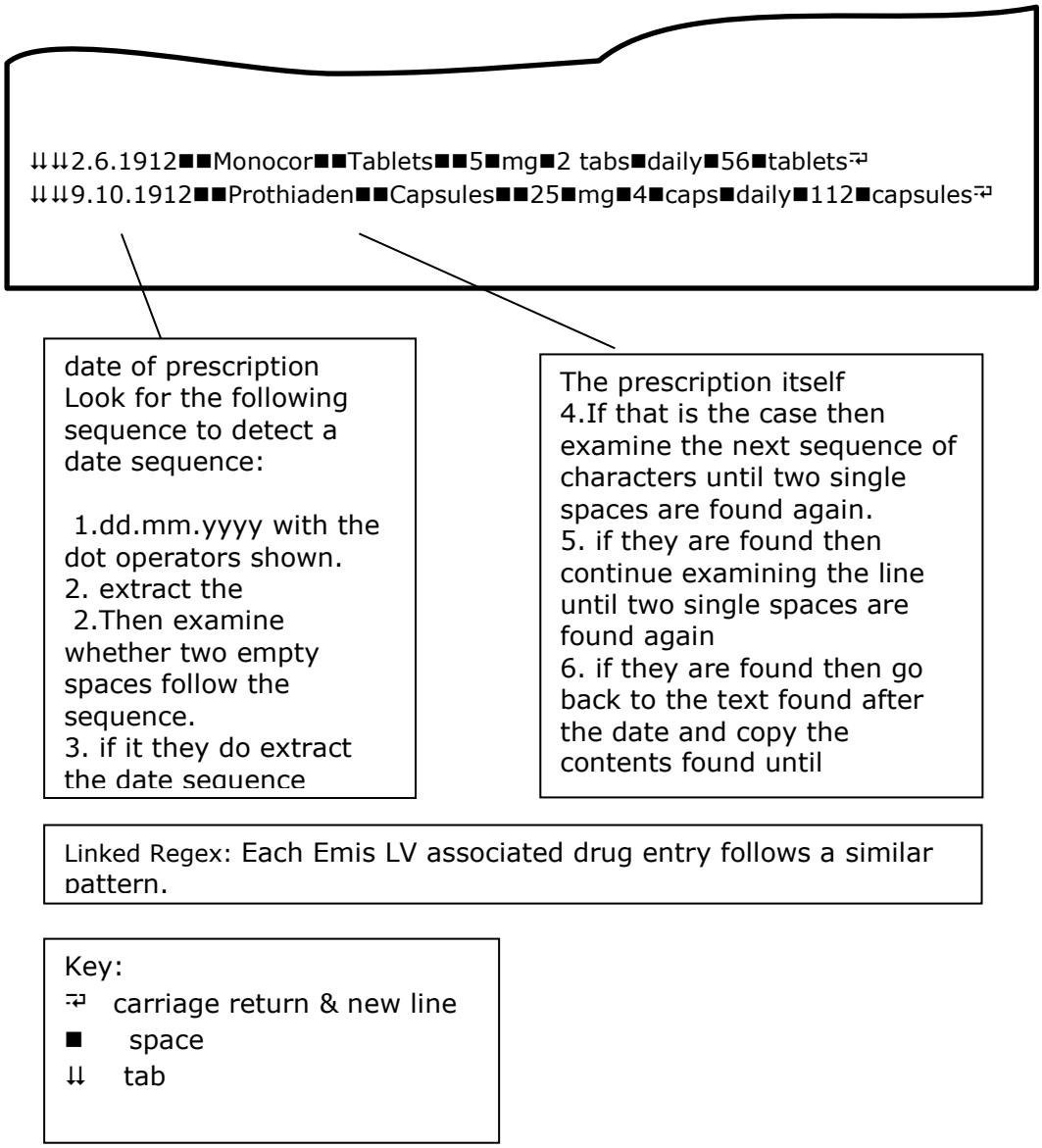


Figure 33 – Emis LV: extraction of prescription date, name, dose, form and quantity

4.6.3.1 Extracting Drugs Information from Emis LV

This refers to extracting the drugs that the patients have been prescribed during consultations.

Regex to implement the above

@"(?<drugDate>\d{1,2}\.\d{1,2}\.\d{2,4})\s{2}(?<drug>(?<name>.+?)\s{2}(?<form>.+?)\s{2}(?<dose>.*)(?=\s\w+))";

Rationale

- The structure of data is such that data is segmented or delimited by double spaces so it ensures a consistent structure to examine.

- Even though this pattern does allow breakdown of name, dose, form it was considered best to keep the dose and form filters in place and run the data through the filters in case there is the potential that the users may input the data into the wrong fields due to end users they may be variation in spaces due to external corruption like personal input as a result the entire contents of line are also extracted after the date field is extracted in case comparisons of the joined up version of name, dose and form are necessary.

4.7 Producing the Emis PCS Data Files

4.7.1 Introduction to Emis PCS data extraction

Emis PCS is part of Egton Medical Information Systems Limited, the same company that owns Emis Lv and Emis Web.

4.7.2 Problems encountered when extracting data

4.7.2.1 Data format and layout issues

Software programming issues

There were less software problems encountered at this point in Emis PCS as many were sorted out in other editions and lessons learned. There were no significant software problems for Emis PCS other than checks already being done such as variable names and accuracy of regexes and ensuring C# code was consistent with previous modules that make up the Extraction Software.

4.7.3 Downloading the Emis PCS files

See appendix 8.9 for download instructions.

4.7.4 Splitting the Emis PCS file into subsections

In the context of an Emis PCS record, the plain text record is composed of tab separated values going in a linear fashion from left to right with data occupying straight lines rather than carriage returns used. The following are the subsections identified and what they cover:

- Patient record – This covers the patient details that will produce the patient demographics for each section.
- Investigation – This covers test results, measurements, letters with dates for each.
- Diary entries –This covers around two-thirds of the rest of the record and covers topics such as GP comments, drugs prescribed, history and examinations.

The following rules apply when identifying the subsections marked above:

The Regexes:

@"(Registration\sDetails)";

@"Investigation\t";

@"(Due\sDiary\sEntries)";

Rationale:

- Due diary entries in a regex is a useful characteristic to use as it also covers overdue diary entries section.
- The diary entries section that is produced contains a mixture of the information that is required from the medical file and that required for the all consultations file. This is because it is grouped together by the record structure and the only way to have a consistent cut-off point was to use the heading for Due Diary Entries. However the data in this section is quite distinguishable and unique whether it's used for Medical or for the All Consultations extractions.

4.7.5 Demographics

These are the demographic variables identified for Emis PCS:

- Name
- Forename
- NHS ID
- DOB
- Address
- Registered GP
- Usual GP
- Gender
- Postcode
- Ethnic origin
- Marital status
- Dispensing

- Trading Partner
- Residential inst.
- Hospital Number
- Records At
- Walking Quarters

4.7.5.1 Name, Gender & Address

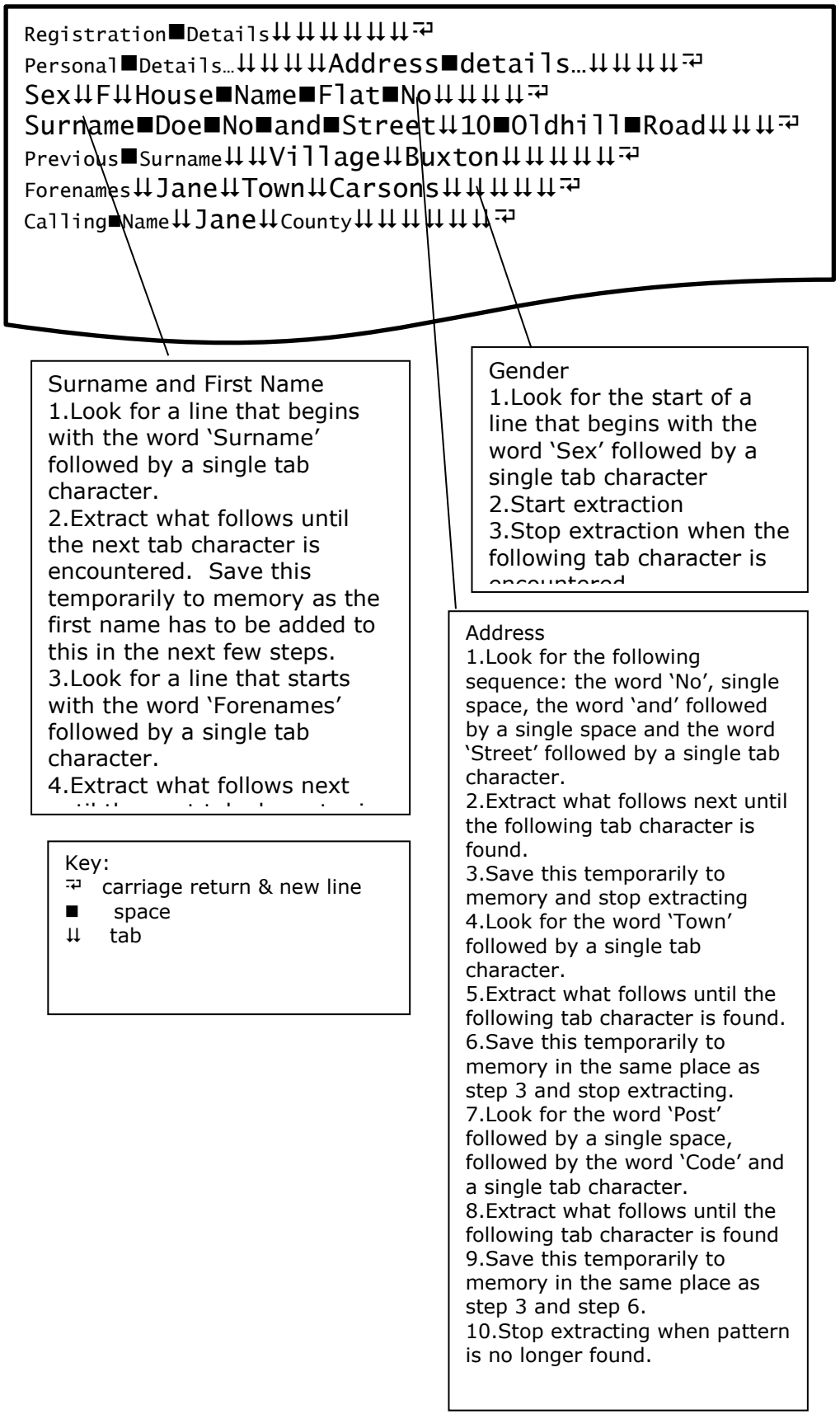


Figure 34 – Emis PCS: extraction of name, gender and address

4.7.5.2 Name (Forename & Surname)

This refers to the first name and surname of the patient.

Regex:

@"(?<surname>(?<=^Surname\t)(?<name>.*?(\t))|((?<=^Forenames\t))(?<name>.*?(\t)))";

Rationale

- Both the surname and first name begin the lines in question and as a result need to be separately extracted but saved in the same 'capture' when it comes to storing what has been extracted. This will ensure that the two bits of information do not become separated from each other during the extraction process and can be accessed together afterwards. This is why the word 'name' occurs twice in the coupled regex expression for surname and first name.

4.7.5.3 Address

This refers to the residential address of the patient.

Regex

@"(?<=No\sand\sStreet\t)(?<address>.*?(\t))|(?<=Town\t)(?<address>.*?(\t))|(?<=Post\sCode\t)(?<address>.*?(\t))";

Rationale

- The address data item presents an interesting challenge due to the nature of the record structure. The record structure has an emphasis on data that is linear so that ordinarily a data item may have a carriage return and a continuation of the data item this is not the case in an Emis PCS record which has data items on their own line in their own demarked areas with continuation using

carriage returns. The effect of this is that each line starts off with a new data item different to the one above.

- The address straddles several lines interrupted by other unrelated data items such as DOB, surname and first name. As a result it was necessary to couple several regexes together and use a mechanism in the regex that allowed temporary storage in the same memory area so all the parts extracted could then be accessed together when the time came to write out the extractions to an external file.

4.7.5.4 Gender

This refers to the sex of the patient.

Regex

```
@"(?<=^Sex\t)(?<gender>.*?(\t))";
```

Rationale

- In Emis PCS records the gender is given as a single character such as F for female and M for male. However the approach should be to extract all the potential data contents that are found as this could change in future editions or even be entered at the practice level as Male and Female.

- In the interests of best practice the ability to extract gender based on tab characters was thought to be better design.

4.7.5.5 Date of Birth, Postcode and Marital Status

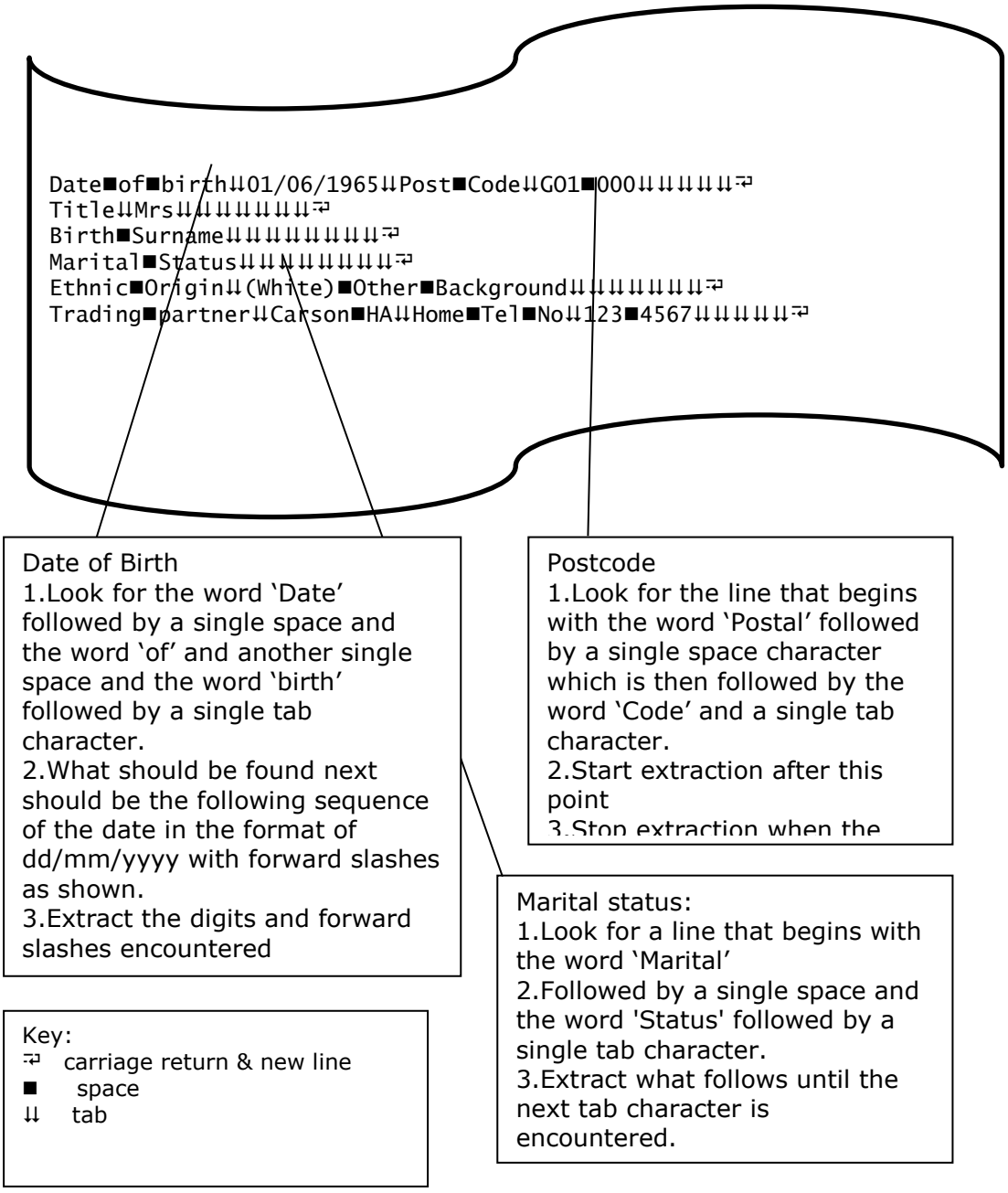


Figure 35 – Emis PCS: extraction of date of birth postcode and marital status

4.7.5.6 Date of Birth (DOB)

This refers to the date that the patient was born.

Regex

@"(?<=^Date\sof\sbirth)(?<dob>\t\d{1,2}\/\d{1,2}\/\d{2,4})";

Rationale

- There are other date values which are present so it's important to make use of the unique label associated with the data item to ensure a consistent and accurate extraction.
- The DOB data item also shares space with the postcode value so it's important that extraction is done only for the digits that are required.
- For the remote possibility that the date value could change from dd/mm/yyyy to dd/mm/yy the regex expression will be able to handle either of these scenarios that occur after the label identifier.

4.7.5.7 Postcode

This refers to the postal code that the patient has their residence.

Regex

@"(?<=Post\sCode\t)(?<postcode>.*?(\t))";

Rationale

- Although in the interests of consistency with other platforms in the earlier regex of address this postcode value is included and is extracted again here and classed as a separate variable because of the way it is treated in an Emis PCS record.

4.7.5.8 Marital Status

The marital status data item is indicated with its own label within the record and this is used to identify it in a unique and consistent manner. The following rules apply when locating the marital status:

Marital status is extracted using the following regular expression:

```
@"(?<=^Marital\sStatus\t)(?<maritalStatus>.*?(\t))";
```

Rationale:

- Marital status can have different choices in potential data items such as married, divorced, single, remarried or separated.
- The definition list may be updated to changes by GP software systems or even in legal terms; as a result it is much more robust to extract the contents that occur between the two labels.

4.7.5.9 Number, Registered GP and Usual GP

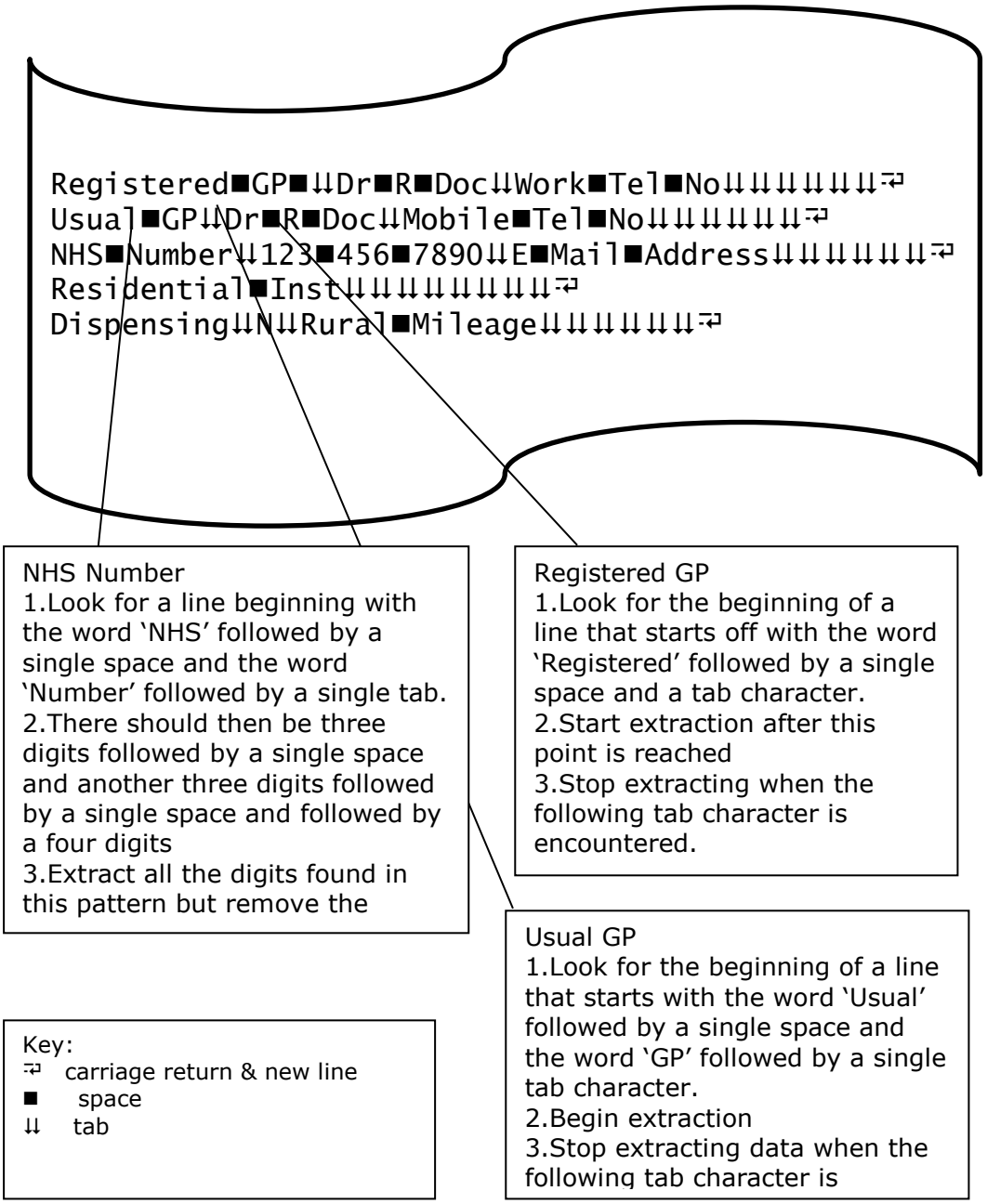


Figure 36 – Emis PCS: extraction of nhsID, registered GP and usual GP

4.7.5.10 NHS Number

This is the unique ten digit identifier for the patient as defined by the National Health Service.

Regex

@"(?<=^NHS\sNumber)(?<nhsId>\t\d{3}\s\d{3}\s\d{4})";

Rationale

- The NHS number has its own label that uniquely identifies the data itself however the NHS ID itself is separated by a sequence of empty spaces; this can also be used in combination with the label when recognising its location and unique appearance.

4.7.5.11 Registered GP

Regex to implement above:

@"(?<=^Registered\sGP\t)(?<registeredGp>.*?(\t))";

Rationale

- This method is robust to accurately to handle any future changes that this regex all the data it could possibly encounter in the narrow sphere given to this data item.

4.7.5.12 Usual GP

This refers to usual general practitioner that the individual may see for their appointments at the practice.

Regex

@"(?<=^Usual\sGP\t)(?<usualGP>.*?(\t))";

Rationale

- At best this is one of the few data items that can be classed as being subjective in its interpretation. The term 'usual' can mean different things to different patients and it's quite possible that they may have several 'usual GPs' so it is not recommended to try and develop a regex that would be tied down to identifying an individual GP.

4.7.5.13Ethnic Origin and Trading Partner

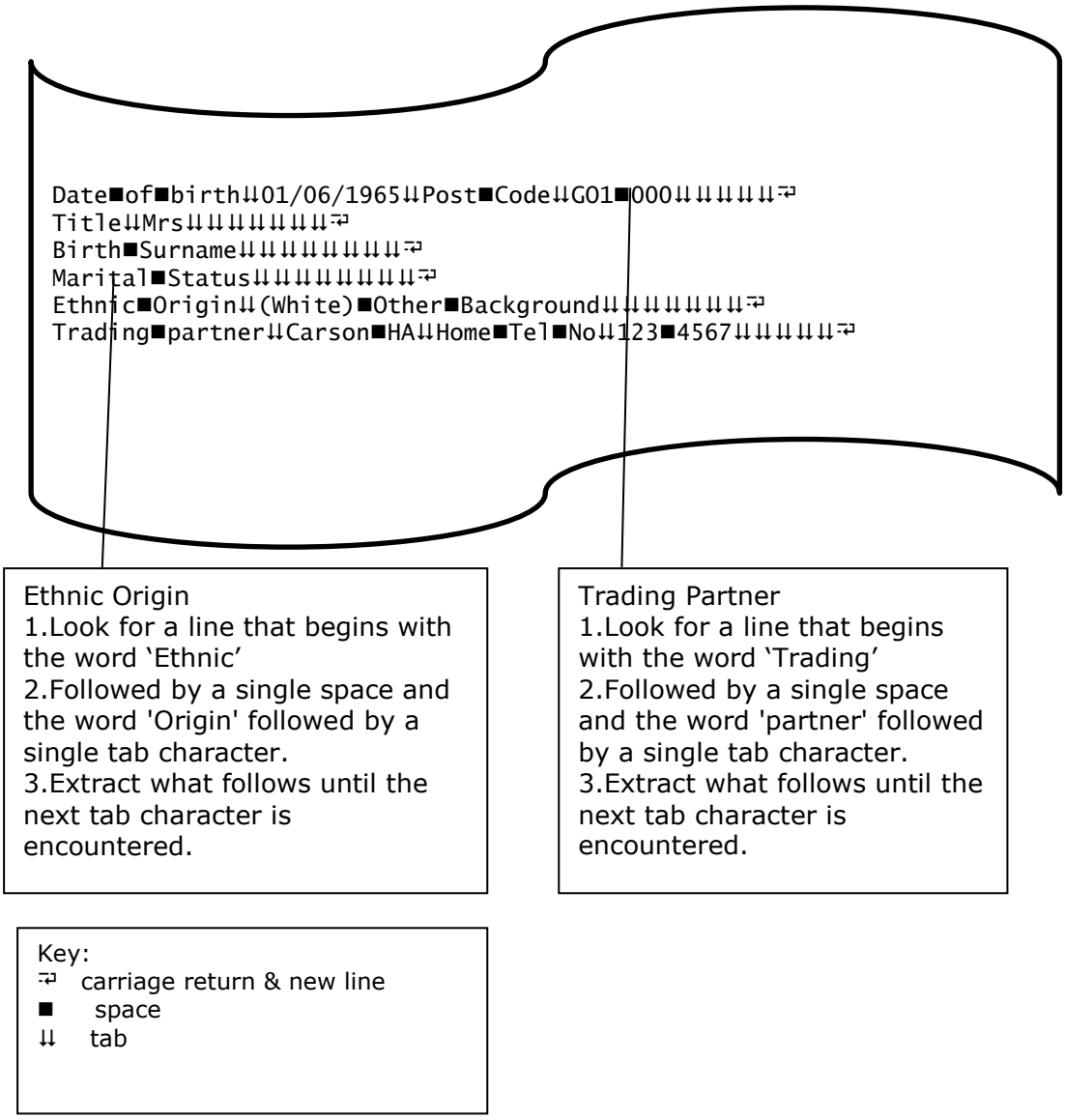


Figure 37 – Emis PCS: extraction of ethnic origin and trading partner

4.7.5.14Ethnic Origin

This data item has its own unique label attached to it to indicate where it is to be found. As a result it can distinguished easily from any other data item.

Ethnic origin is extracted using the following regular expression:

@"(?<=^Ethnic\sOrigin\t)(?<ethnicOrigin>.*?(\t))";

Rationale:

- For reliability it is much more robust to extract ethnicity value from between the two data labels than create a list from looking at the current values found in the records to allow for differences

4.7.5.15Trading Partner

Some GP practices list the trading partner in the records and this refers to the practice's financial relationship with the UK health service in the local environment. For example NHS commissioning board.

Trading Partner is extracted using the following regular expression:

@"(?<=^Trading\spartner\t)(?<tradingPartner>.*?(\t))";

Rationale:

- Trading relationships can change and this would need updating in the records. One option implemented was to analyse all the Emis records and use coding to match whichever was in the list of trading partners. However if a practice's trading relationship changes it's possible that a custom list may not include the name. It made much more sense to use the label as that is least likely to be modified compared to the actual trading partner.

4.7.5.16Dispensing and Residential Inst.

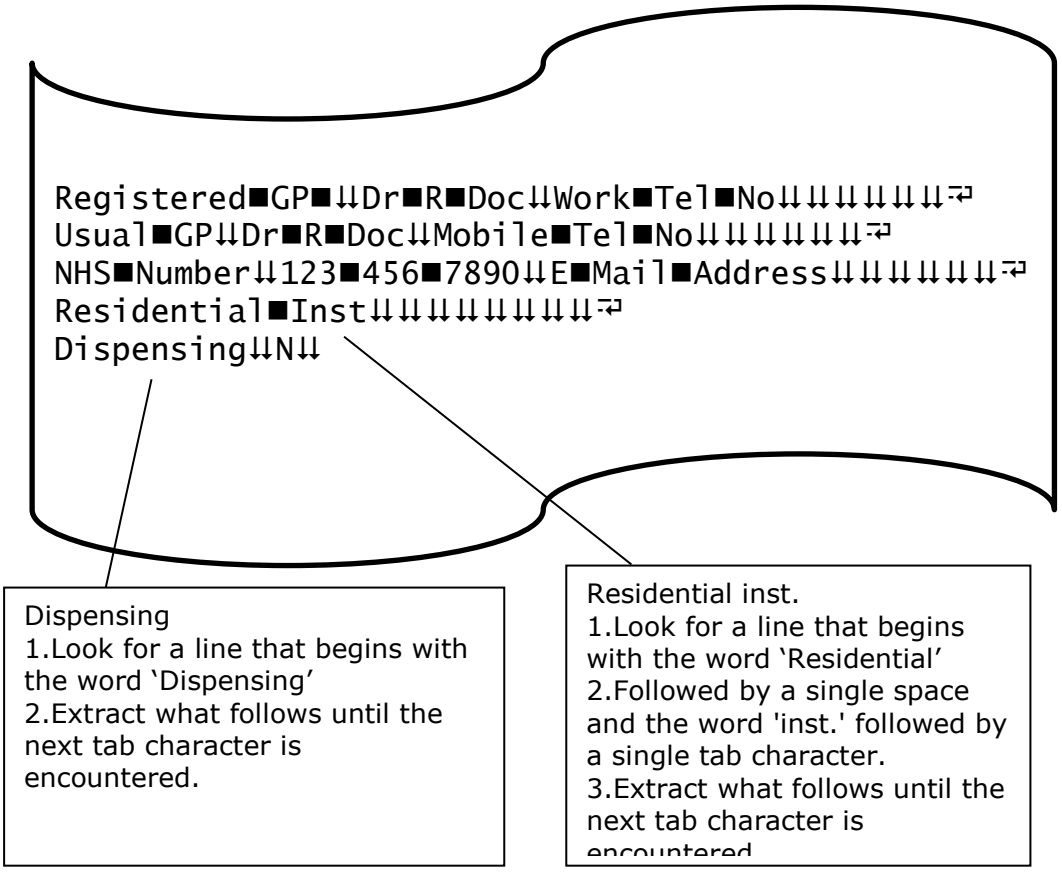


Figure 38 – Emis PCS: extraction of dispensing and residential instruction

4.7.5.17 Dispensing

This information is regarding whether the practice is a dispensing one. Again not directly related to the patient demographic details but may provide interesting information.

Dispensing is extracted using the following regular expression:

@"(?<=^Dispensing\t)(?<dispensing>.*?(\t))";

Rationale:

- Some practices may introduce this service or discontinue and although the response to this can be range of options to yes or no or left blank as has been seen in the majority of records the safest and long lasting approach is to use the identifier label that precedes where the value would ordinarily reside.

4.7.5.18Residential Inst.

- This refers to the patient's instruction regarding their residential data. Again like the RPP Road miles, Walking quarters the Residential Inst. has been found to be empty in all Emis PCS records so far.

- It makes it difficult to ascertain the use of these fields in Emis PCS records but nonetheless it's important for accuracy purpose that an attempt to extract should it exist.

The variable Residential inst. is extracted using the following regular expression:

@"(?<=^Residential\sInst\t)(?<residentialInst>.*?(\t))";

Rationale:

- Without observing real-life examples the specifics of Residential Inst. are unknown as to what it is composed of and looks like. However it maybe that patients may indicate some instructions in this field at some point. Therefore makes sense to build that into the extraction process in case.

4.7.5.19Hospital Number, Records At & Walking Quarters

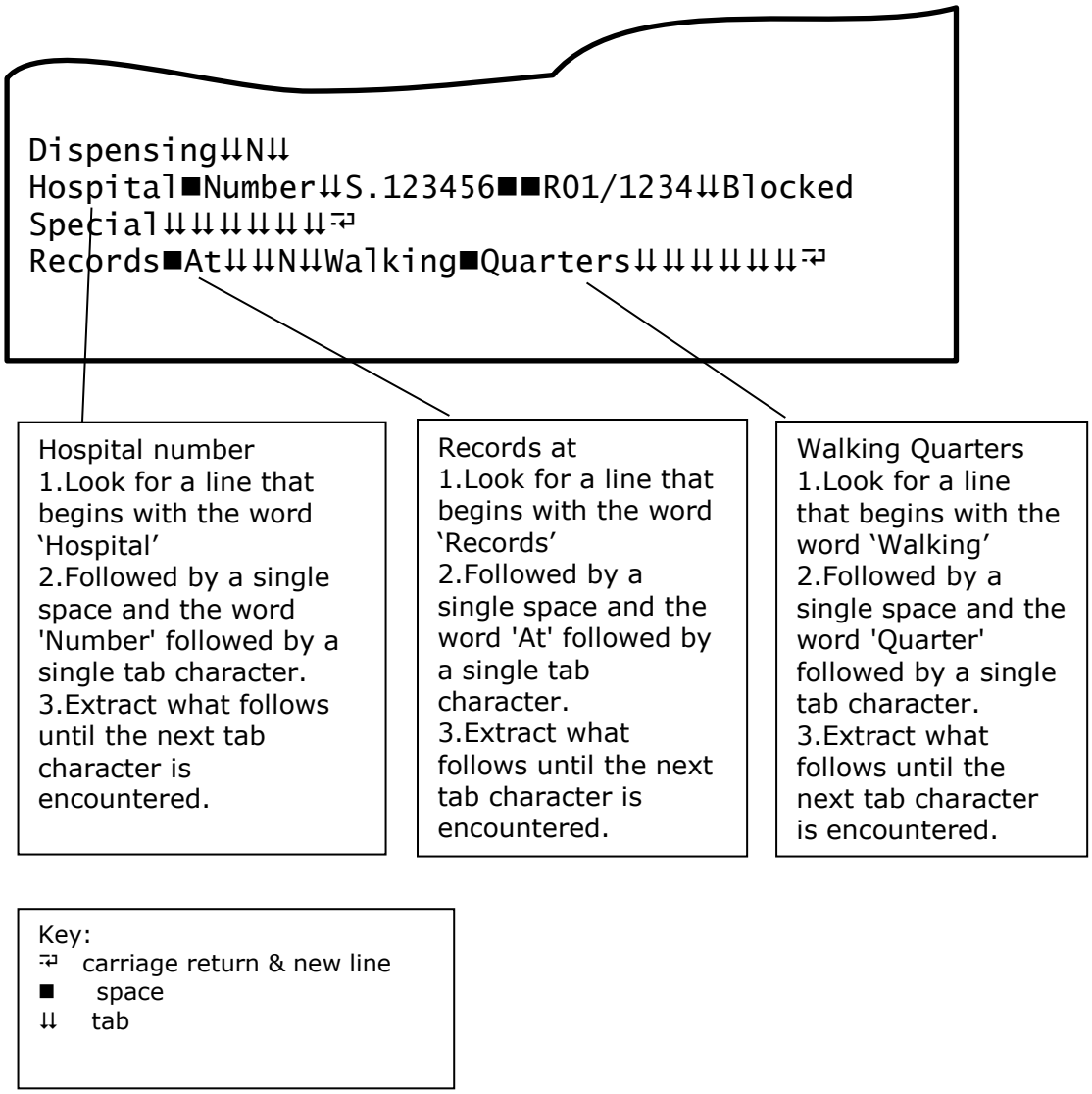


Figure 39 – Emis PCS: extraction of hospital number, records at and walking quarters

4.7.5.20 Hospital Number

This refers to the hospital number of the patient. The NHS ID is still the universal identifier for health care in the UK but if a hospital number is present it will be extracted.

Hospital Number is extracted using the following regular expression:

```
@"(?<=^Hospital\sNumber\t)(?<hospitalNo>.*?(\t))";
```

Rationale:

- Although the length and formatting of the hospital number means a regex can be designed if these things are modified they could conflict with the NhsId. To avoid this occurring the label that precedes this variable will do the work of highlighting the position to the regex.

4.7.5.21 Records At

This refers to where the GP records of the Patient are kept. It also refers to where the paper records are stored.

Records At is extracted using the following regular expression:

```
@"(?<=^Records\sAt\t)(?<recordsAt>.*?(\t))";
```

Rationale:

- In some cases a practices can be part of a larger group, some in different places of the same town and paper records may be stored at a larger one for example. As a result greater accuracy lies in using the label rather than writing code that can match the practice name to the one being sought in the record. It is possible to do it in that manner but if there are minor variations in names being matched a false error may occur.

4.7.5.22 Walking Quarters

This refers to a walking quarters measurement associated with patient. It is a measurement label found in all Emis LV records encountered and has been found to be empty thus far. It will however have a regex to handle any items if they are ever present. Walking quarters like RPP/Road miles gives an idea of distance measurements but since it is has always been empty thus far it cannot be said for sure what metric it is applied with.

Walking Quarters is extracted using the following regular expression:

@"(?<=Walking\sQuarters\t)(?<walkingQuarters>.*?(\t))";

Rationale:

- Even though it has been seen to be empty in case Emis LV records not knowing anything about what possible metrics or formatting could be used it is sensible to utilise a regex that can handle the scenarios. It is better for long term functionality that a regex attempt to extract any data values associated.

4.7.6 Medical

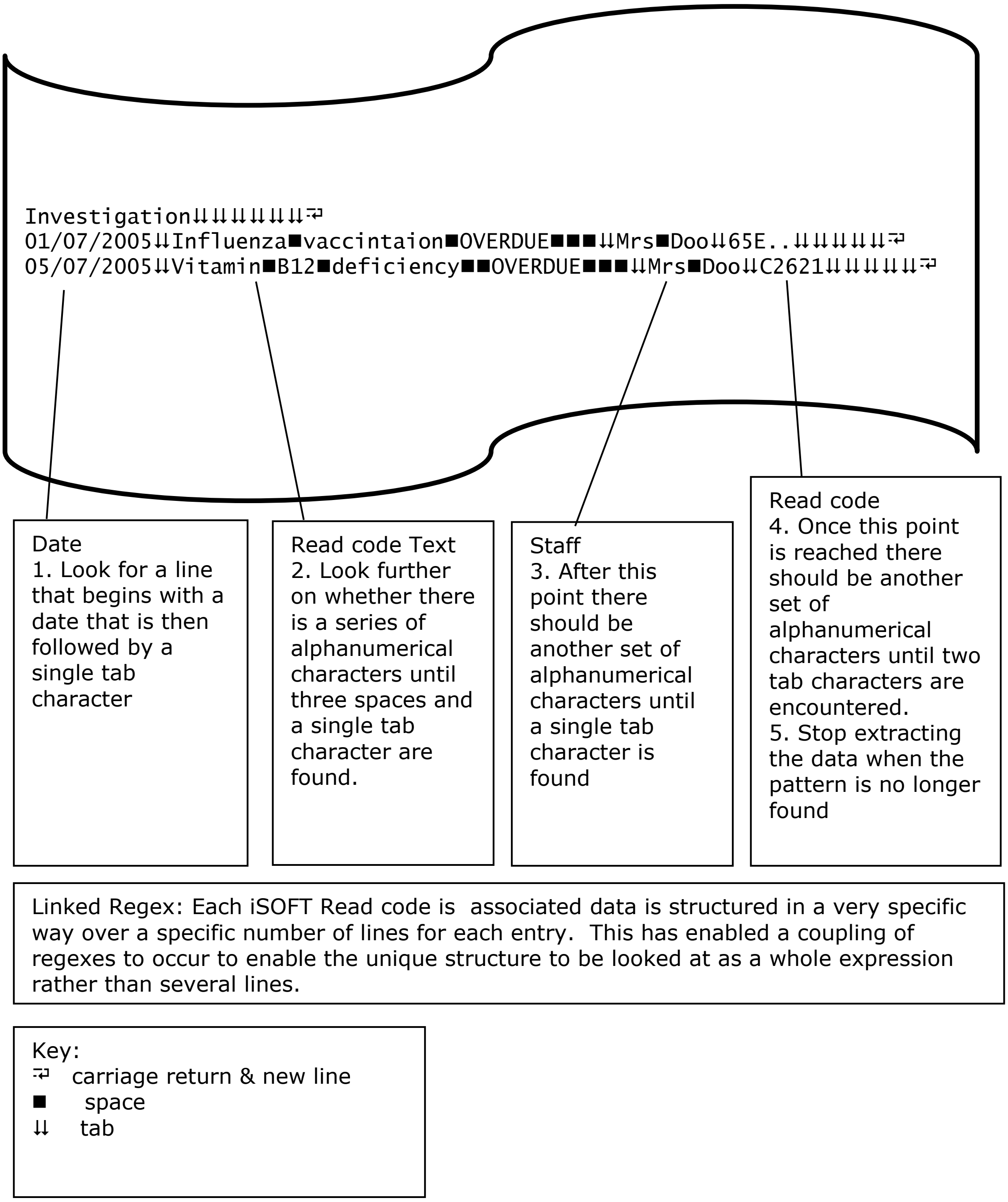


Figure 40 – Emis PCS: medical extraction of date, Read code text, staff and Read code

This refers to extracting the Read codes its associated data to examine the nature of the therapy provided to the patients and looks at the Investigation.txt subsection.

Regexes to implement the above:

@"((?<date>(\d{1,2}\/\d{2}\/\d{4}))\t(?<readcodeText>\w.*?)\s\s\s\t(
?<staff>\w.*?)\t(?<readcode>\w.*?)(?=\t\t))";

Rationale

- Investigation subsection covers many different pieces of patient information such as measurements and test results. However in this case we are looking for information that is encoded using the Read code structure.

- The structure itself from a regex point of view is not easy to isolate because it can have a mixture of alphanumeric characters in varying forms with different cases though always being of 5 such characters in length. From a regex perspective this could be any 5 word character that occurs in this section.

- To ensure robust and accurate isolation of the Read code it's best to extract it in the context of its Read code description and the date on which the entry is made as that adds a uniqueness whilst also extracting linked data to its entry which can help to give context to the Read code itself and be potentially more useful.

4.7.7 Checking of Read or Clinical Codes

An extra step of validation of extracted Read codes was used in the same manner as described in systmOne. This was achieved by matching Read codes against a dictionary list of terms provided by the Health and Social Care Information Centre (HSCIC) so that the corresponding Read codes could be checked as being valid (see appendix 8.10 for more information).

4.7.8 Drugs EMR section for Emis PCS

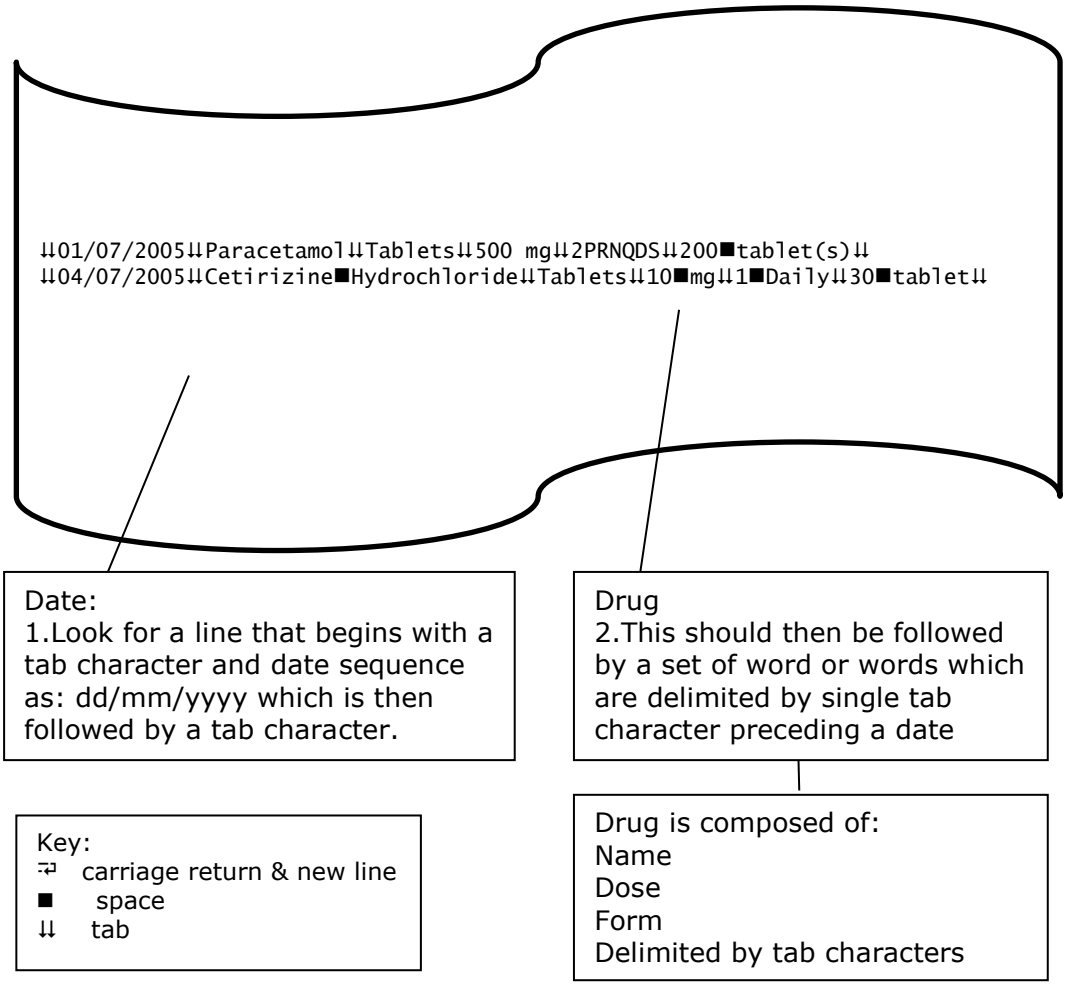


Figure 41 – Emis PCS: drug extraction of date, drug comprising of name, dose and form

This refers to extracting the drugs that the patients have been prescribed during consultations. It looks at analysing the DiaryEntries.txt subsection of the Emis PCS record.

Regex to implement the above:

```
@"(?<drugDate>^(\d{1,2}\/\d{2}\/\d{4}))\t(?<drug>(?<name>\w.*?)\
s\s(?<form>\w.*?)\s{1,2}(?<instruction>\w.*?)(?=\t(\d{1,2}\/\d{2}\/\d
{4})))";
```

Rationale

- This section of diary entries has lot going on within its own subsection as also has a lot of unstructured text and headers which vary in occurrence so their order cannot be used as the basis of anything consistent.

- The best approach was found that the area of diary entries that deals with drugs prescribed always began with a date field. It was also found that the line itself was structured so that the name of drug followed the date; delimited by two spaces and the form followed thereafter again delimited by two spaces and what follows is the dosage and instruction for that particular prescription.

- As names of drugs are mentioned in unstructured areas of this section it was considered that the best approach was to use a combination of the preceding date, delimited spaces and tab characters so that only drugs in the context of the day they were prescribed were extracted.

- However the unstructured areas of diary entries still contains a wealth of health information as a result a separate approach and regexes were adopted so that this could also be utilised. This is covered in the All Consultations section (Emis PCS) later on in this chapter.

4.8 Producing the Emis Web Data Files

Emis Web is a GP software version from Egton Medical Information Systems Limited. This section covers the process of extracting the patient, medical and drugs files from Emis Web records.

4.8.1 Problems encountered when extracting data

4.8.1.1 Data format and layout issues

One issue encountered was that the dates in Emis Web EMRs were separated by hyphens instead of forward slashes but were in the correct format of dd-mm-yyyy and code was written to deal with this. The iterative and incremental process of evolving the regex definition lists so that forms and dosages were extracted effectively continued for Emis Web (appendix 8.8)

4.8.1.2 Software programming issues

One of the problems encountered for Emis Web EMRs was that during testing a situation was found where no NHS identifiers were being extracted. Upon checking with EMRs the error was found where the C# code was not correctly developed to match the NHS identifier pattern. The C# code was changed and the NHS identifiers were correctly extracted for all Emis Web EMRs.

4.8.2 Downloading the Emis Web files

See appendix 8.9 for download instructions.

4.8.3 Splitting the Emis Web file into subsections

Once the records have been received it was necessary to examine their unique structure. In this study there are three editions of Emis products and each one of them is unique in structure. As a result it was necessary to treat each edition as its own standalone product for the purposes of this study. Emis Web records are most similar to Emis LV as its closest relation as there are crossover consistencies. The first step was to mark out the general structure of the record in terms of headers to get a general idea of

how information was being held where the information regarding demographics, medical and drug files could be extracted from. A high-level view from the outset is useful in getting a good idea of where data is being repeated and which section has the data as a full set and which section is it being summarised within.

4.8.4 Marking the Boundaries of an Emis Web Record

The boundaries within the record are defined by the positions that each respective identified header is located. This defines the beginning and ending of what is used in the regex expression that splits the record into subsections. Each subsection then has this header as its main title and this subsection is then used further on the software. The need for subsections is so that there can be a specific context to the nature of the data and an added uniqueness to the way that the data can be expressed as a regex and a more organised approach to separating the record into its constituent parts. The rules below are explain how these subsections are defined for Emis Web records and how these are then expressed as regexes to define the edges of each subsection so that the whole record splits out into the accurate sections.

4.8.5 The Main Emis Web Files Extracted

In the context of an Emis PCS record, the record is composed of tab separated values going in a linear fashion from left to right with data occupying straight lines rather than use carriage returns to use up the following lines. Much of the data is not separated by too many headers and different bits of patient data appear one after another without headers separating them. We can only speculate on why there are not many natural headers but this does not stop the process of using splitting as it can be done on the basis of the type of data found if the headers indicating the start and stop of specific sections is not present. The following are the subsections identified and what they cover:

- Patient Record
- Active Problems
- All Medication
- All Allergies
- Health Status
- Immunisations
- Planned Events
- All Consultations
- Attachments

The following rules apply when identifying the subsections marked above:

The Regexes:

Patient Record = @"(\s{39})";

Active Problems = @"(\r\nProblems\r\n)";

All Medication = @"(Medication)";

All Allergies = @"(Allergies)";

Health Status = @"(Health\sStatus)";

Immunisations = @"(Immunisations)";

Planned Events = @"Planned\sEvents";

Consultations = @"(Consultations)";

Values and investigations = @"(Values\sand\sInvestigations\s\(All\))";

Referrals = @"(Referrals)";

Attachments = @"(Attachments)";

Biological Values = @"(Biological\sValues)";

Rationale:

- Emis Web records are consistently structured using these headers even if the headers themselves do not contain any information pertaining to their own sections.

- Using a this stable structure it's possible to split out sectors which are consistent for all Emis web records
- Splitting out sections of a record allow the data within the respective sections to gain an added uniqueness by self-containment and a greater emphasis for that subsection to exist independently of the larger record which has all the information in every subsection put together.
- It ensures that regexes are more likely to be unique and be coded with greater clarity as it's more likely the data being extracted is unique within the specific subsection.

4.8.6 Demographics

These are the demographic variables identified for Emis Web:

- Name
- NHS ID
- Usual GP
- DOB
- Address
- Registration Date
- Dispensing
- Advocacy needs
- Transport needs
- Language
- Patient Type

4.8.6.1 Patient Name, NHS Id, Date of Birth and Address

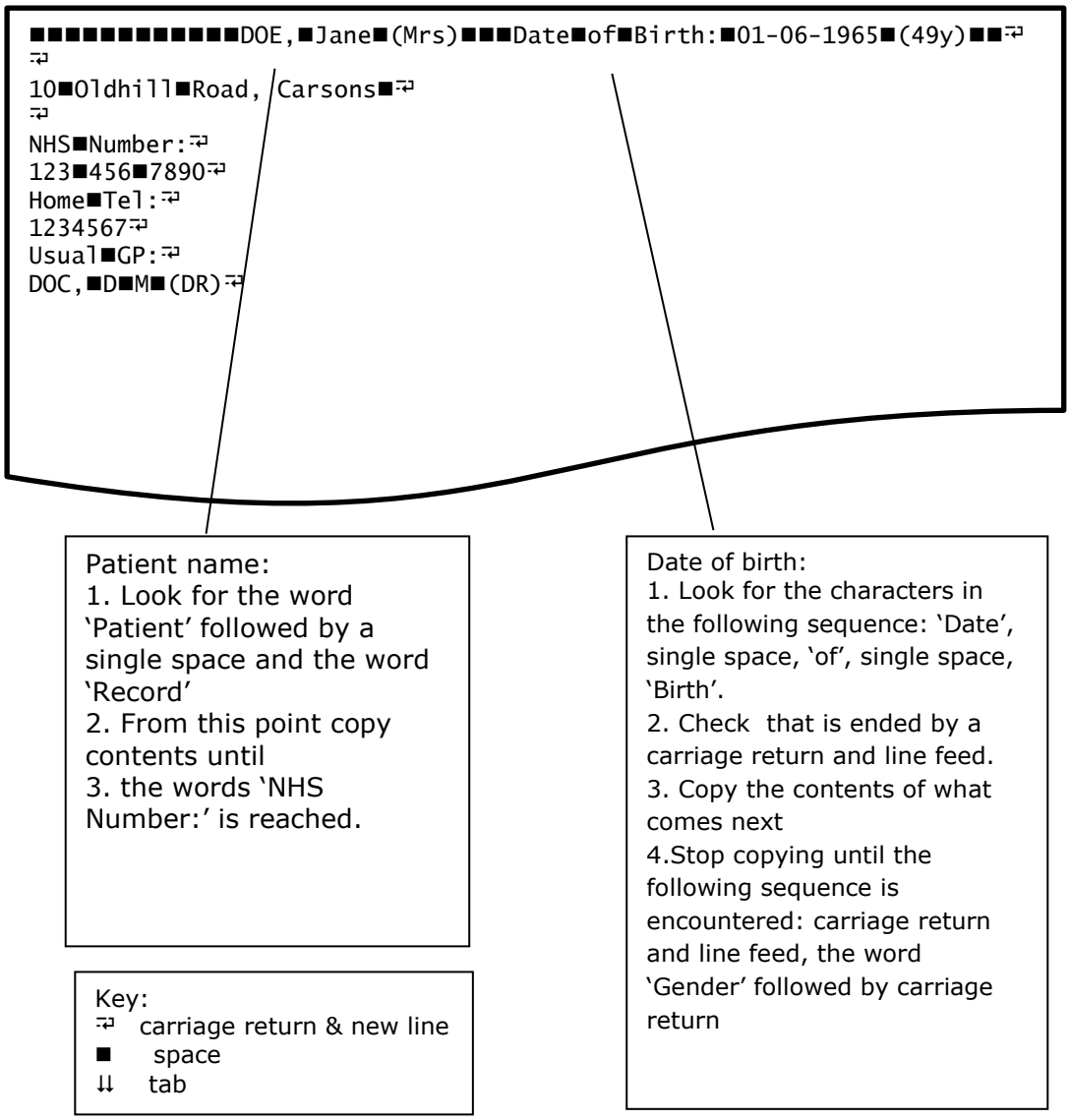


Figure 42 – Emis Web: extraction of patient name and date of birth

4.8.6.2 NHS Id and Address

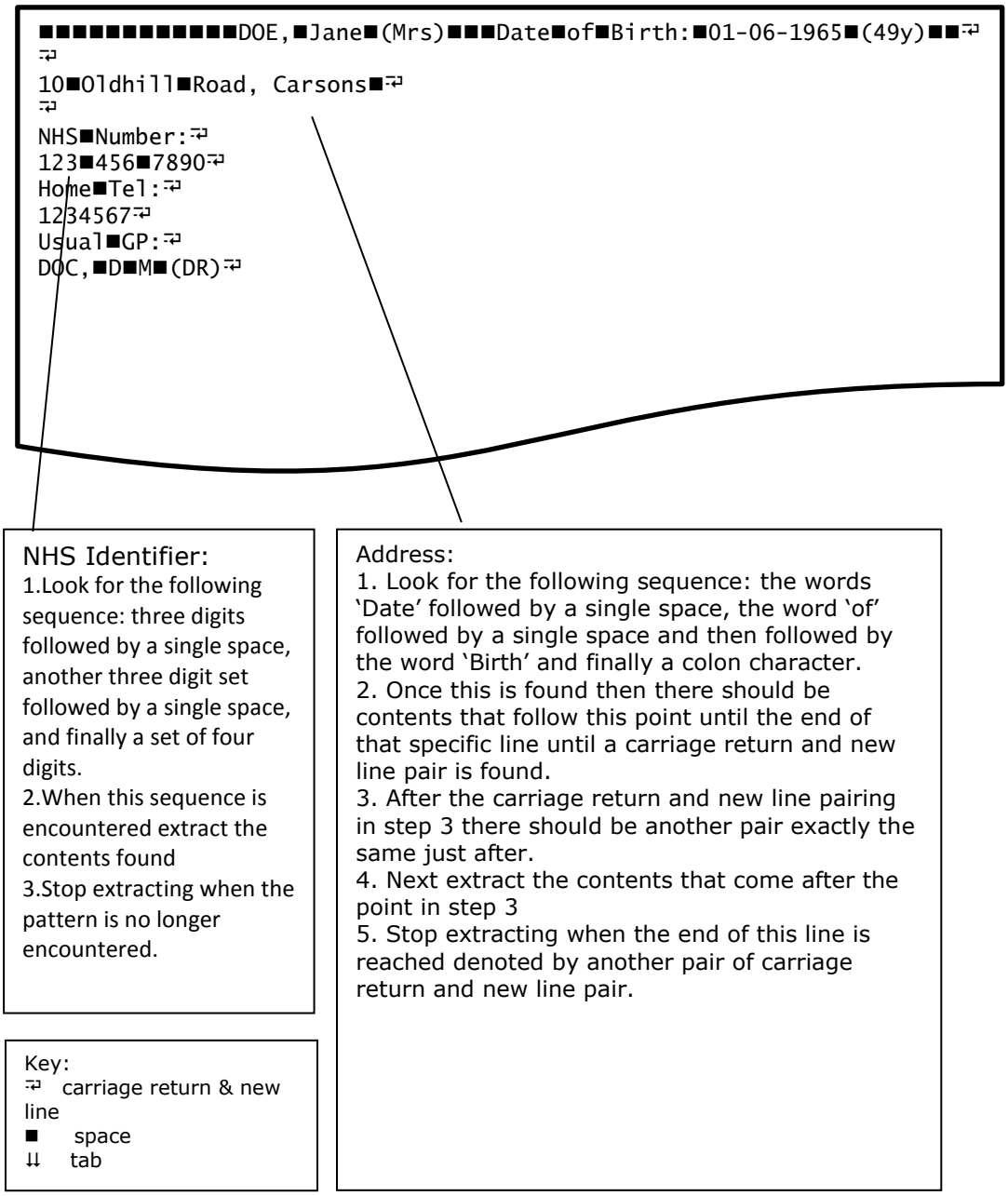


Figure 43 – Emis Web: extraction of nhsid and address

4.8.6.2.1 Name

This refers to the patient first name and surname.

Regex:

@"(?<name>.*(?=Date\sof\sBirth))";

Rationale:

- The name value data item does not have an own label that identifies the specific position.
- However it is the very first data value that is encountered and the next data value does actually have its own label identifying the data value that follows after which is the date of birth data item.
- By using the label associated with date of birth it's possible to express a stable pattern that will consistently and accurately identify the name by first looking for this label and extracting what comes immediately before it and extract the title and full name.

4.8.6.2.2 NHS ID

This refers to the unique identifier as allocated by the NHS.

Regex:

@"(?<nhsId>\d{3}\s\d{3}\s\d{4})";

Rationale:

- The NHS Identifier has an own label indicating where it sits, however the number sequence with spaces is a much more distinctive way of identifying the data itself
- This is because it is the only time in the subsection that this combination of digits and spaces occur.

- Some Emis editions contain a specific Emis ID which is shorter and looks different to the NHS ID so by expressing the NHS ID in terms of its unique structure is a much more robust approach than just using the label which could be subject to change whereas the NHS ID is used nationally and is something Emis will less likely to change dramatically.

4.8.6.2.3 DOB

This refers to the date of birth of the patient.

Regex:

@"(?<dob>(?<=Date\sof\sBirth\:).*(?=\())";

Rationale:

- The date of birth data item is found sandwiched between its label identifying its location and the piece of data that refers to how this translates into age in years.
- For our purposes the date of birth is what we are most interested in and as a result need to express an extraction rule that identifies both aspects to either side of its location.
- By expressing the data item in this way it ensures the whole information is extracted irrespective of the notation used such as dd/mm/yyyy or using dd.mm.yyyy or another approach. With the above approach any notation will be correctly extracted.

4.8.6.2.4 Address

This aspect of data refers to the current residence of the patient.

Regex:

@"((?<=(Date\sof\sBirth\:.*)\r\n(\r\n))(?<Address>.*\r\n))";

Rationale:

- Address does not have its own label associated identifying its location.
- It would not be a sensible approach to express it in terms of a house number followed by a location as this could be mistaken for some other piece of data.
- It does however always consistently occur in the same manner for all Emis web records that is just after the date of birth data item.
- Expressing the extraction rules to occur after the date of birth value is the most robust and consistent pattern for all the records.

4.8.6.3 Usual GP, Patient Type and Registration Date

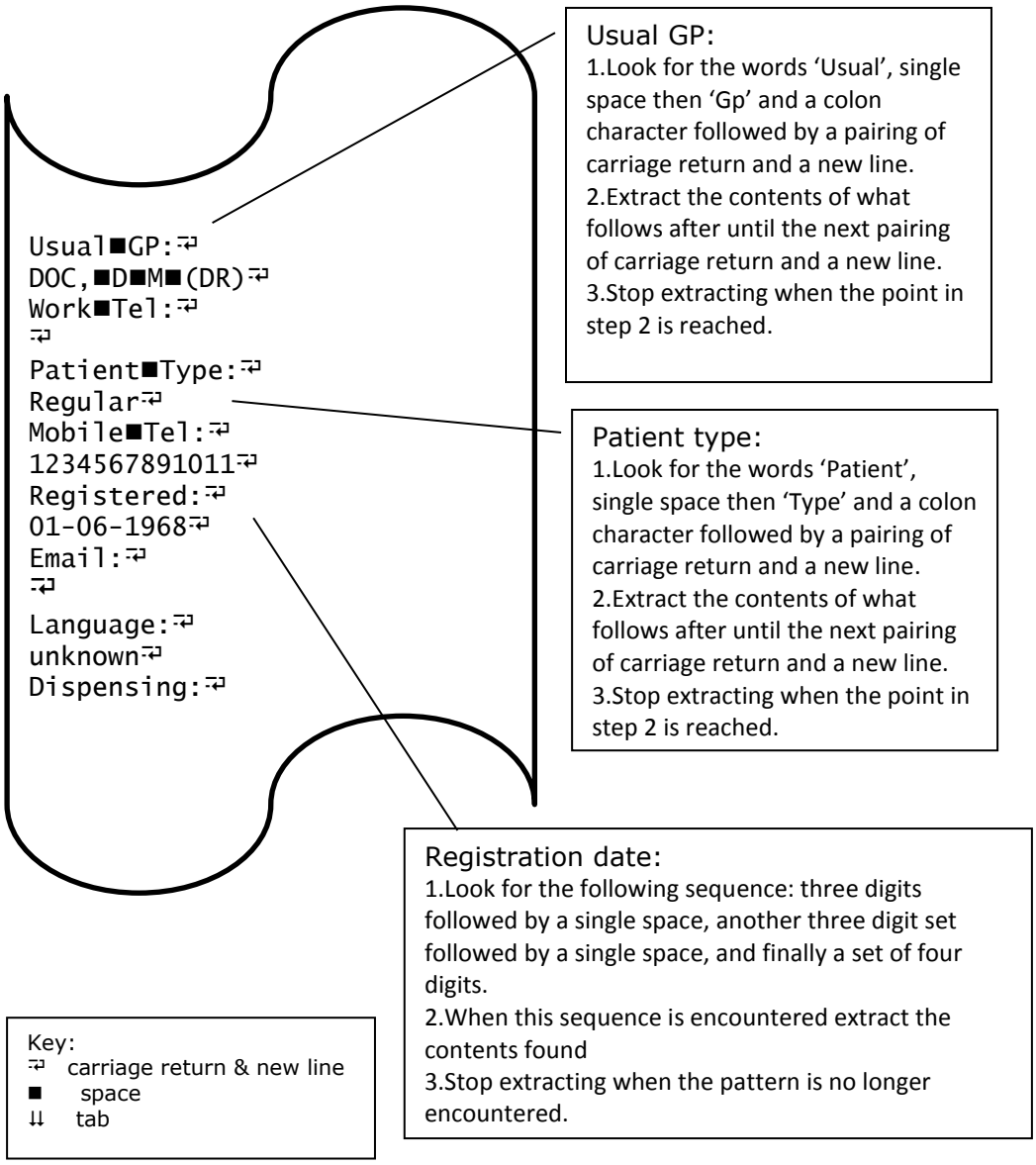


Figure 44 – Emis Web: extraction of registration date, usual GP and patient type

4.8.6.3.1 Registration Date

This refers to the date the patient registered at the practice as a patient.

Rules to identify them:

Regex:

@"(?<=Registered\:.*\r)(?<registrationDate>.*?\r)";

Rationale:

- This data item does have an own label denoting its location.
- However this occurs in the next line and this has to be accounted for in the expression so that the label itself is not extracted but what occurs immediately is extracted.
- The label itself seems to be stable part of Emis Web records in that consistently occurs in every record and is highly likely to be populated as every patient will have a registration date to be a patient at the practice in the first place.

4.8.6.3.2 Usual GP

This data item refers to the name of the general practitioner that the patient sees on a regular basis during consultations.

Regex:

@"(?<=Usual\sGP\:.*\r)(?<usualGP>.*?\r)";

Rationale:

- The label in this context is on the line above and as a result it is necessary to use the carriage return and new line to detect the location of the actual data item which is the usual general practitioner information.

- In other GP systems this value can be different such as an actual medical group for example. As a result it not appropriate to denote the words 'Dr' as part of the extraction rule even though the data item always starts with this approach.
- It is much more accurate to extract the contents that occur after the label in case the 'Dr' part may be missing.

4.8.6.3.3 Patient Type

This refers to the general status of the patient whether they are a temporary patient or a regular type patient permanently registered at the practice.

Regex:

@"(?<=Patient\sType\:.*\r)(?<patientType>.*?\r)";

Rationale:

- This data item has its own label associated with its location with data being on the next line following.
- The type of patient can be viewed as a slightly ambiguous use of wording which could even vary among practices. Loosely it's referring to whether the patient is a permanent patient or a temporary patient; however in many Emis Web records this information is recorded as being 'Regular'.
- This is can be misinterpreted as to whether this refers to a the volume of appointments that the patient is recorded with
- As a result the best approach was to express the rules so that all the contents referring to the label where extracted.

4.8.6.4 Language and Dispensing

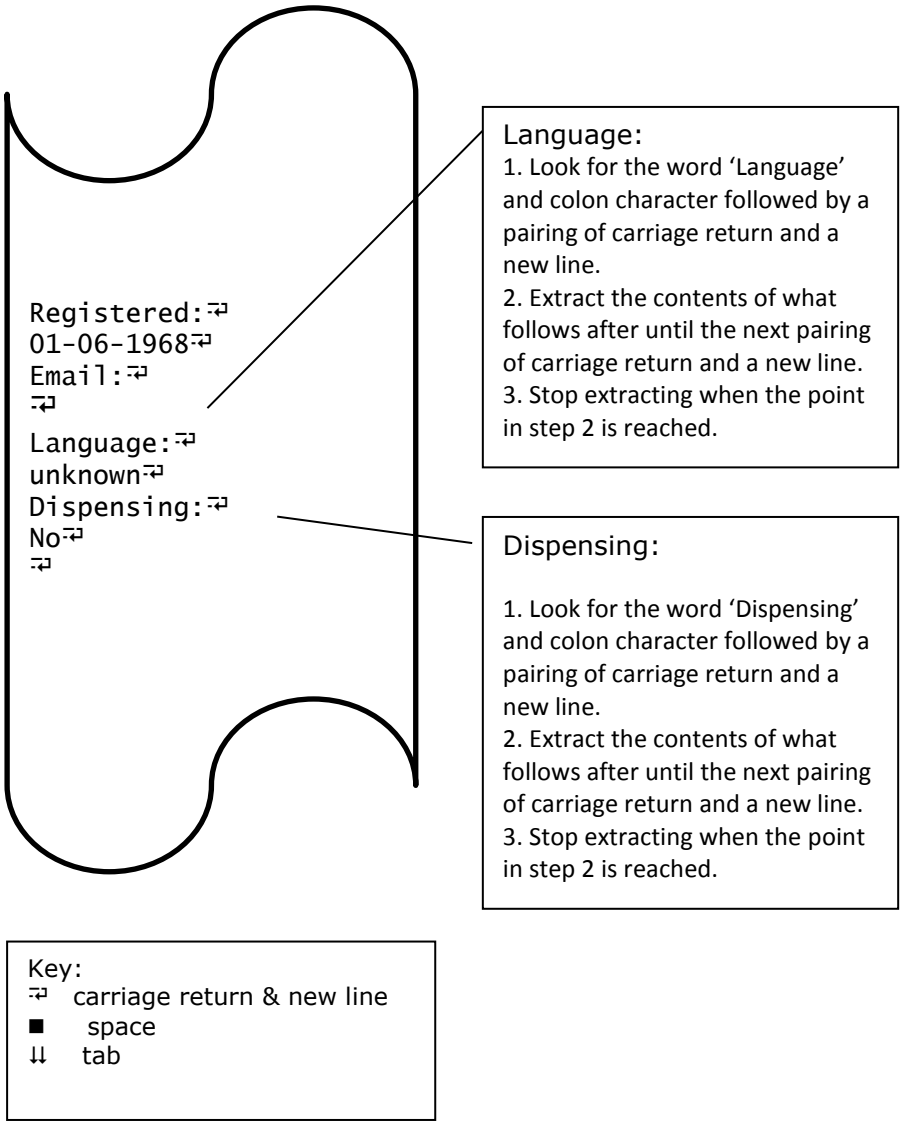


Figure 45 – Emis Web: extraction of language and dispensing

4.8.6.4.1 Language

This refers to the main language that the current speaks.

Rules to identify them

Regex:

@"(?<=Language\:.*\r)(?<language>.*?\r)";

Rationale:

- Language has its label associated with it in terms of where the data item is to be found.
- The language data itself is found underneath the label on the next line which is the reason for using carriage returns and new lines within the expression so that the label is used but not itself extracted.
- The language notation itself uses standard English dictionary terminology however it would not be a sensible approach to use a list of possible languages as many records in Emis Web actually record the 'unknown' for this data item. Whether that is true or whether that is a default value until update occurs is not clear.
- As a result it is safer to use the label rather than capture a list of possible language options.

4.8.6.4.2 Dispensing

This information is regarding whether the practice is a dispensing one. Again not directly related to the patient demographic details but may provide interesting information.

Dispensing is extracted using the following regular expression:

```
@"(?<=Dispensing\:.*\r)(?<dispensing>.*?\r)";
```

Rationale

- Some practices may introduce this service or discontinue and although the response to this can be range of options to yes or no or left blank as has been seen in the majority of records the safest and long lasting approach is to use the identifier label that precedes where the value would ordinarily reside.

4.8.6.5 Advocacy and Transport Needs

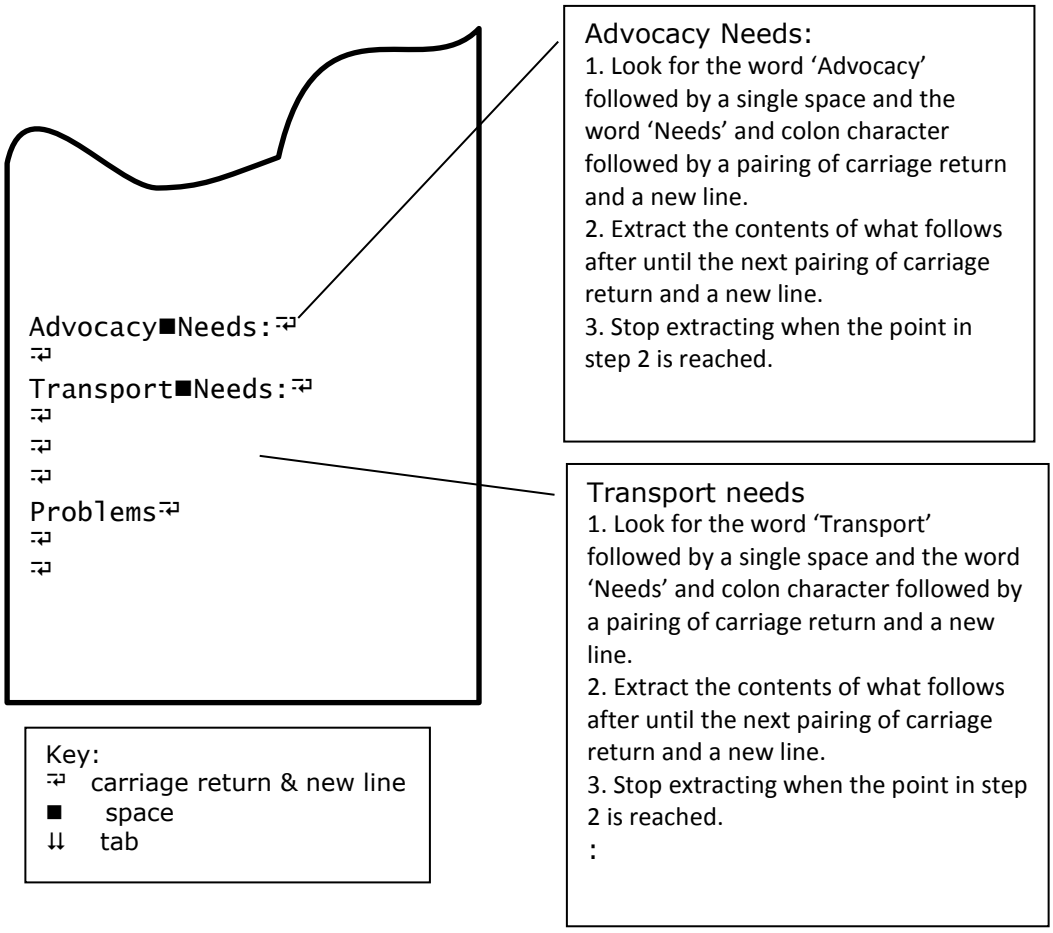


Figure 46 – Emis Web: extraction of advocacy and transport needs

4.8.6.5.1 Advocacy Needs

This information is about whether they need a chaperone in the same room during any consultation or assessment. Advocacy needs are extracted using the following regular expression:

@"(?<=Advocacy\sNeeds\:.*\r)(?<advocacy>.*?\r)";

Rationale

- This data item has its own label associated with its location with data being on the next line following.
- All of the Emis Web records encountered have no information on advocacy needs so it is not possible to evaluate the formatting or

specific structure. As a result the best approach was to use the associated label and extract the contents as advocacy needs could be similar to free text length.

4.8.6.5.2 Transport Needs

This information indicates whether they need any type of assistance to get to the GP practice. Transport needs are extracted using the following regular expression:

```
@"(?<=Transport\sNeeds\:.*\r)(?<transportNeeds>.*?\r)";
```

Rationale

- This data item has its own label associated with its location with data being on the next line following. Though all the Emis Web records encountered so far don't have any data available so that the structure or length could be evaluated. As a result the safest option is to use the label name and carriage returns to ensure accurate extraction if data is present and it varies over time.

4.8.7 Medical

Medical data revolves around attempting to extract the nature of the therapy received be examining the encoding used during consultations.

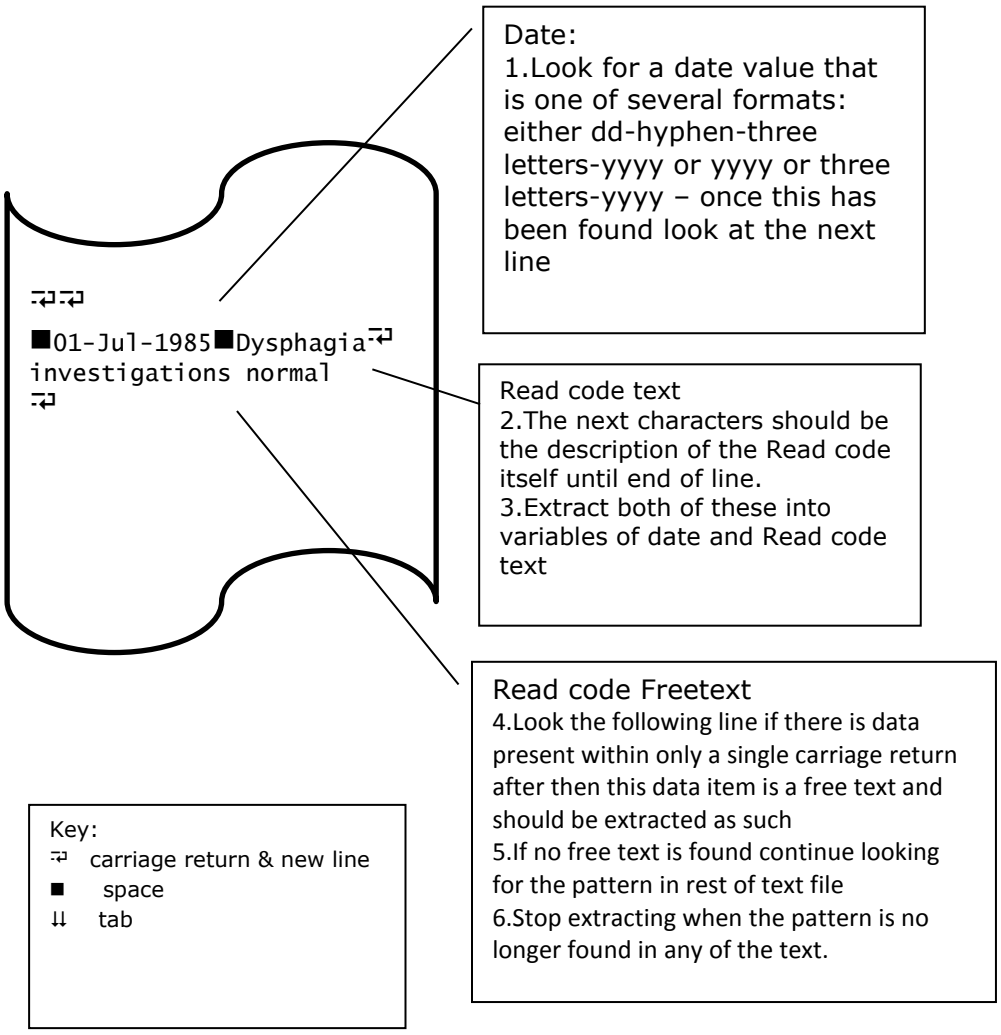


Figure 47 – Emis Web: Extraction of medical subsection of date, Read code text and Read code free text

In Emis Web records this information is stored within the active problems subsection.

- Date – this is the day month and year recorded for the consultation
- Read code text – Medical data encoded using the description about what the specific problem is all about.

- Read code free text – There may be occasions where in addition to the description of specific problem diagnosed the practice may record supplementary information that may be important or relevant.

Regex:

@"((?<=(?<date>(\d{2,4}\-\w\w\w\-\d{2,4}|\b\d{4}|\w\w\w\-\d\d\d\d)))\s(?<readcodeText>(.*?)))\r\n?(?<readcodeFreeText>.*?)\r";

Rationale:

- The entries for medical data occur in a consistent pattern beginning with a date to denote a new entry
- Free text associated with the entry is sometimes present so an optional part of the regex accounts for its presence or absence.
- Using the association of the Read codes with their respective dates can help in further analysis by using the dates to examine the history of the patients.
- Using dates uniquely identifies the start and finish of an entry if the entry may be recording the same Read code. It helps to distinguish Read code entries.

4.8.7.1 Advanced Matching for Read codes

Emis Web records contain rich information to the Read code descriptions but not the Read codes themselves. This is also the case with Emis LV records though is not the case with the Emis PCS edition of GP software. To gain a complete picture of the Read codes present it's necessary to pattern match the descriptions received from the above expression against a dictionary list of known terms and their respective Read codes (141).

When a Read code description matches a term the corresponding Read code is returned to be associated with the description found. If the Read code description cannot be matched then it forms part of the free text data automatically for manual review. Consistent with method used for Emis LV.

4.8.8 Drugs

This file provides an extraction of the drugs that the patients have been prescribed.

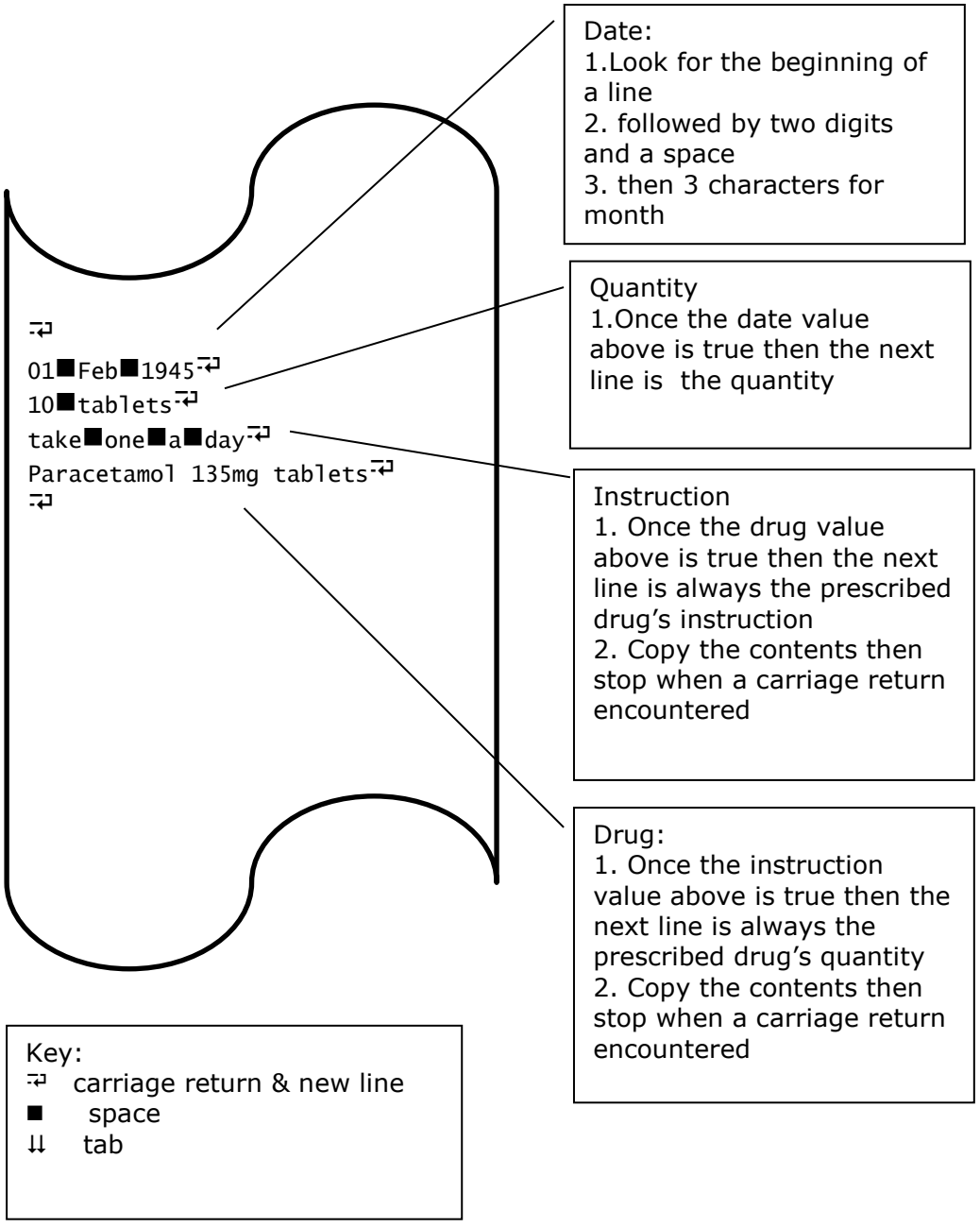


Figure 48 - Emis Web: extraction of prescriptions looking at date, quantity, instruction and drug information of name, dose and form

4.8.8.1 Emis Web Drugs

The drug prescriptions are captured in the All Medication subsection of Emis Web records. The following section outlines how prescription information is extracted.

- Date – the actual day, month and year that the prescription was recorded on the GP record.
- Quantity – If any quantities regarding the prescriptions are indicated they are recorded in the GP record.
- Instruction – This refers to how the patient is told to consume/apply the medication(s) issued. It normally is also found labelled on actual dispensed medication.
- Drug – the name of the drug, dosage and form are recorded as well.

Regex:

@"(?<drugDate>(\d{1,2}\-\w{3}\-\d{4}))(\r)(?<qty>([^\r])*)(\r)(?<instruction>([^\r])*)(\r)(?<drug>([^\r])*)(\r)";

Rationale:

- Using the date as a starting point is using a naturally accurate starting point as each entry appears in this manner in the actual record. It can thus be used to denote the start and end of a specific prescription.
- Isolating the prescriptions according to the date that is recorded for them allows encapsulation of each entry from similar other entries

and ensures the extraction is able to handle even the same drug prescription but where the date is different.

- Using the pattern of carriage returns means the regex can effectively be coupled to extract all the grouped information for that specific prescription such as instruction and quantity and form.

4.9 Other variations noticed between the different GP software systems

Another finding was that there was variation between the different GP software systems about where key information covering patient, medical and prescription data was found and in the general layout of each GP system's respective EMRs. SystmOne records had the most consistent layout in terms of labelling of sections. For example language and ethnic origin labels being present in the EMR even if the data around those labels had not yet been inserted. This was different to iSOFT records where if a section did not have any data then its labelling would not present either. Emis LV, Emis Web and Emis PCS did not follow this concept and their layout was not subject to change in a manner like systmOne's design. Understandably, each GP system has produced their respective EMRs with a unique design. In all the GP systems demographics encountered were usually found at the top of the EMRs. Prescriptions and medical information was generally found grouped in their own areas and not intertwined with one another or mixed up. This was a common theme among the EMRs that the data sections containing a type of information were distinctly found together. Differences in the user interface were apparent when GP practices were visited. This has a natural inherent difference in the EMRs layout upon extract at the GP level. Yet these differences were have been accounted for by each module in the extraction

software that pertains to a specific GP systems. As such the extraction software is composed of five modules.

4.10 Post extraction: how the files were put together from the different GP systems

By looking at the structure of the EMRs encountered the Extraction Software detects which GP system the EMR is derived from and then chooses the most appropriate module to trigger so that the series of regexes can be applied to complete the extraction process from EMRs into extracted files. Once the modules are completed depending on which EMRs are encountered a final set of unified files are produced ready to be analysed. This is shown in Figure 49:

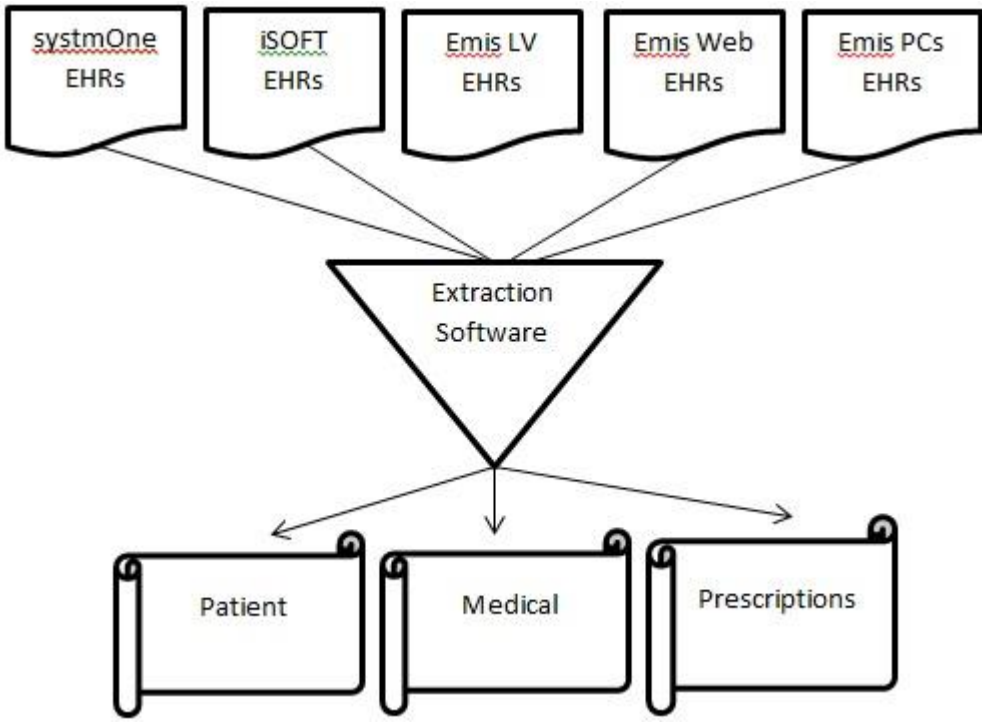


Figure 49 Extraction Software (ES) interaction with EMRs collected from GP practices and final rendering into 3 unified files covering demographics, medical and prescriptions extracted from all EMRs

4.11 Strengths and Limitations

Strengths and Limitations of the approach taken have already been covered in 2.11. However there are some further points gained from the current chapter that warrant discussion. There is a lot of obvious and subtle differences in the nature of the EMR data from the different GP software systems encountered. There were different names for the similar sections of an EMR. For example one GP system would call it immunisations and another would term the same section as vaccinations. In some cases such as iSOFT if there are no immunisations recorded then the EMR would not have that section at all. In the rest of the GP systems that section would just contain a title. A weakness of using EMRs for data extraction is that they are highly dependent on many factors which can introduce less reliable data such as default computer settings, practice staff entry habits and training, design decisions of GP companies, and disparate naming conventions. These are weaknesses because it is difficult to detect without some sort of first-hand experience of how EMRs are used at the practice level. For example experience of visiting practices and talking to staff and doctors gave an insight into possible factors. One practice reported that they had no updated their Read codes for over a year. In another practice reported during data collection that GP prefers to continue using paper notes to write down the consultation but uses the GP system to record the medicine prescribed and makes a recording of the consultation on the computer as well but handwritten notes are never entered. Another practice complained about the high number of letters that do not add much value constantly being sent between practice and secondary care and how often only important ones are ever scanned onto their GP system. Yet another practice stated verbally that since they had changed over from a different GP system a staff member had not fully

finished entering problems into the new system manually and would only concentrate on entering the most serious retrospective problems since they were also short of staff. As a result of these differences it seems that it would be useful to include a qualitative even ethnographical approach when extracting EMR data so that subtle differences can be recorded and possibly adjusted for in analysis. A feature of using EMRs for data extraction and analysis is that they have been built for use at primary care practices and not research but a strength is that since they are downloadable in an electronic format still they have enormous potential to be used and accurately extracted despite the differences which has been demonstrated by the Extraction Software. The methods used in this study's approach are based on those used in previous studies as well which have been discussed (see 2.12) are less susceptible to these issues as the objective was to extract all EMR data into the Gedling database rather than extract a certain illness as a result it is an appropriate approach to take bearing in mind all the different factors that can influence EMR recordings if a study was looking for something specific. Narrower focused studies using EMRs should seek to supplement their data extraction and verification methods with practical experiences of what software other raw repository they are extracting from in terms of software systems, design and include data about how systems are used by the people concerned to account for differences beforehand.

4.12 The Next Steps: using the EMR derived files in further analysis

The overall objective of the extraction software was to convert all data stored in EMRs into a standardised format similar to THIN tables. However it is not targeted at particular outcome or set of exposures. The aim of processing all EMRs in this manner is to enable future researchers to carry

out studies of any condition recorded in original EMRs. Using the extracted EMR derived data will be an important phase and is presented in the next chapters. It represents the completion of purpose for which the Extraction Software was developed and to illustrate a real world problem solving abilities of using EMR data at the GP level which would otherwise be left unused for a majority of time.

4.13 Summary

A computer program built on the use of regular expressions was designed to extract the relevant data from the EMRs collected which resulted in three output files called patient, medical and drugs. The Extraction Software then joins the files from each GP software system together to create the same three files but now in unified manner across all GP software systems. Patient deals with demographic data, medical lists diagnoses and ongoing therapy and drugs show the medicines indicated. The diverse EMRs are have a structure characterized by the how they have been designed by their respective GP software systems. It creates uniqueness in an EMR's structure which regular expressions are able to handle and take advantage so each expression targets data accurately depending on the EMRs the computer program encounters. By using this approach a mixture of different GP software systems were processed all at once and the unified files produced ready for analysis. This computer program was designed and written by the researcher. Where issues around software programming were encountered these have been listed by GP software system.

5 Comparison of the Extracted Data in the Gedling Database across the different GP Software Systems

5.1 Aims of this Chapter

- Describe and present results of using extracted EMR data from all the GP software systems encountered during practice data collection
- Examine the difference in the GP software systems encountered at practices using the following attributes:
    - Number of Coded/Uncoded text (free text)
    - Prescriptions
    - Consultations

5.2 Introduction

This chapter looks at the EMR data extracted from the Gedling cohort at the GP practice level. It begins by looking at the rationale to examine the EMR characteristics between the different GP software systems and specifically focuses on the amount of free text in records, followed by prescriptions and consultations across the cohort's EMRs. Although there is research on using primary care databases and the differences between them, there is however very little available on the differences between different GP software vendors and their respectively designed EMRs. There is a fundamental missing link in not knowing a comparison between different GP software systems and their EMRs. The objectives of chapter 6 is to look at specific aspects that are common across the EMRs in a manner to allow a like for like comparison to assess the differences and provide a possible explanation for any differences and what the implications may be.

5.3 Background

A significant feature of primary care databases such as THIN, CPRD and QResearch is that they are extracted from the same brand of GP software such as Vision practices for THIN and CPRD and Emis practices contributing to the QResearch database. Therein exists potential that there may be differences including data definition, coding differences, completeness and non-random missing data resulting in measured and unobserved confounders. Studies comparing between research databases have produced mixed results showing similarities and differences. For example one study that compared database differences in population characteristics and in data definitions, recording, quality and completeness used both CPRD and Qresearch to examine whether differences in the databases changed their results. That study concluded that whilst recording differences between the Vision and EMIS computer software systems could result in differential coding of some comorbidities results from both databases were close to each other (143). In contrast another study comparing two databases found a complex pattern of similarities and differences between databases of CPRD and Doctors' Independent Network (DIN) with overall treatment effect estimates not statistically different but the two studies led to different conclusions regarding the safety of beta-blockers and in some stratified groups the effects differed significantly (144). This shows that EMR recordings can vary for a number of reasons; possibilities could include the type of GP software system in use, design of the software interface and practice staff habits and size of practice and patient demography. There may also be the possibility that recording in EMRs is now very much focused towards incentivised national priorities such as QOF. For example rates of electronic recording of smoking status and cessation advice increased highly during QOF's introduction (145).

Less research is available on data recording on health measurements considered less urgent whether that is in terms of recording as often or as consistently in comparison. Another area where documentation of a patient medical record can vary is information recorded in the free text. For example a study found a large amount of extra information relating to diagnoses was stored in the free text of EMRs (146). Therefore further investigation is necessary to examine if EMR differences have the potential to produce misleading and non-generalizable conclusions. So far no studies have been encountered where a cohort's electronic records from different GP software systems have been extracted into a new database to examine the potential differences that may exist between EMRs at the practice level. This novel approach will examine if there are any actual differences by looking at the EMRs in a newly created research database that has EMRs from the five different GP software systems encountered at practices during EMR data collection.

5.3.1 Measuring Key Characteristics of EMRs for Comparison

An empirical comparison of EMRs from different GP systems contributing data to its own cohort-derived primary care database is missing and has not yet been encountered in any research thus far. For the purposes of a fair comparison it is necessary to identify variables that we would expect to be common to all systems and recorded on a regular frequency in a manner that we would be able to make a direct comparison. This will mean that the variables identified should be able to tell us something fundamental in the way the different EMRs have been used at the practice level. Since one of the primary uses of EMRs occurs during the patient-GP interaction and recording this process is one of EMRs most key uses it is sensible to look at what occurs during this interaction. The EMRs are therefore most basically used for recording diagnoses, any other relevant information deemed necessary during patient-GP interaction and if needed

issuing a prescription to treat the problem(s) found. Consequently, this process of patient-GP interaction can be quantified and so measured using the following variables:

- Free text – Text entered by hand and or computer and initiated by either GP or staff related action.
- Prescriptions – The number of drug recordings for that patient
- Consultations – The number of times the patient visits the doctor

5.3 Methods

5.3.1 Study Population

The population was formed from the Gedling cohort patients whom we were able to collect at the EMRs at the practices contacted and permitted data collection. Even though patient consent was received some practices did not permit data collection for any reason they deemed correct. There was no way to counter practice dissent which has been covered in chapter three already. The total number of participants was 717. A full breakdown of the cohort population is provided in chapter 3 please see section: the Gedling population participating in EMR data extraction for the current study.

5.3.2 Extracting age, sex, socioeconomic status and other factors

Sex and date of birth was extracted from medical records using regular expression patterns. Definitions of how the remaining factors were calculated are given below:

Age

Age was calculated by using the extracted date of birth and comparing this to the collection date of the EMR. This was calculated for all participants who were then categorised into the following age groups:

- youngest-29,
- 30 < 49
- 50 < 65

- 66 < 75
- those over 76 years of age

Sex

Sex of participant was extracted from the collected GP records where this item was present inside the records.

Socioeconomic status

Deprivation measures were taken from previous collected data during the most recent Gedling follow-ups in 1991, 2000 and 2007. Using measures of deprivation creates a single measure of how deprived an area is using post-code linked characteristics in the Lower Super Output Area (LSOA) postcode data set from the Office of National Statistics (ONS) (147). These can be used to calculate a Townsend score as a single measure of deprivation. For the purpose of these analyses Townsend scores were divided into quintiles where 5 was the most deprived and 1 the least deprived quintile. This was achieved by linking the postcode value of the participants to the LSOA-linked Indices of Multiple Deprivation 2007 (IMD) data from the ONS.

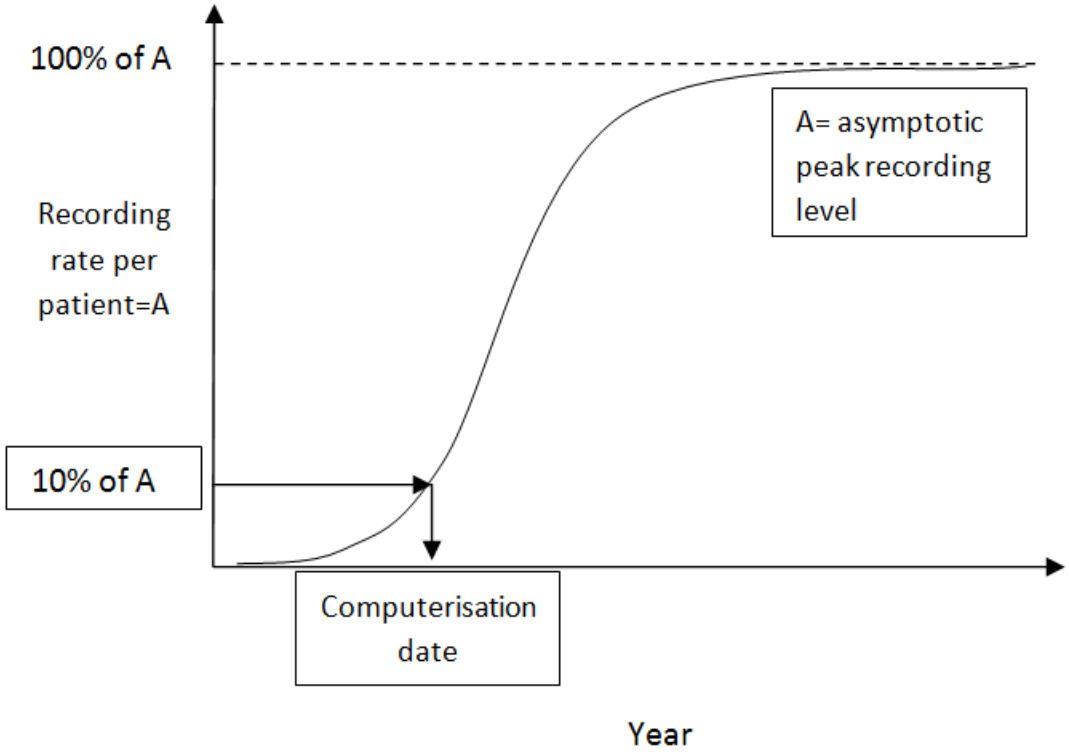
Social Class

Social class categories were taken from previous collected data from previous Gedling follow-up studies in 1991, 2000 and 2007. There were no modifications made to these definitions. Social class was defined by occupation according to the Registrar General's classification (148). They had been defined as social class 1-5 with 1 being the highest social class denoting professional occupation. These were achieved by merging the data extracted from the collected GP records to the data sets from previous follow-ups using their National Health Service (NHS) unique identifier.

5.3.3 Computerisation Date

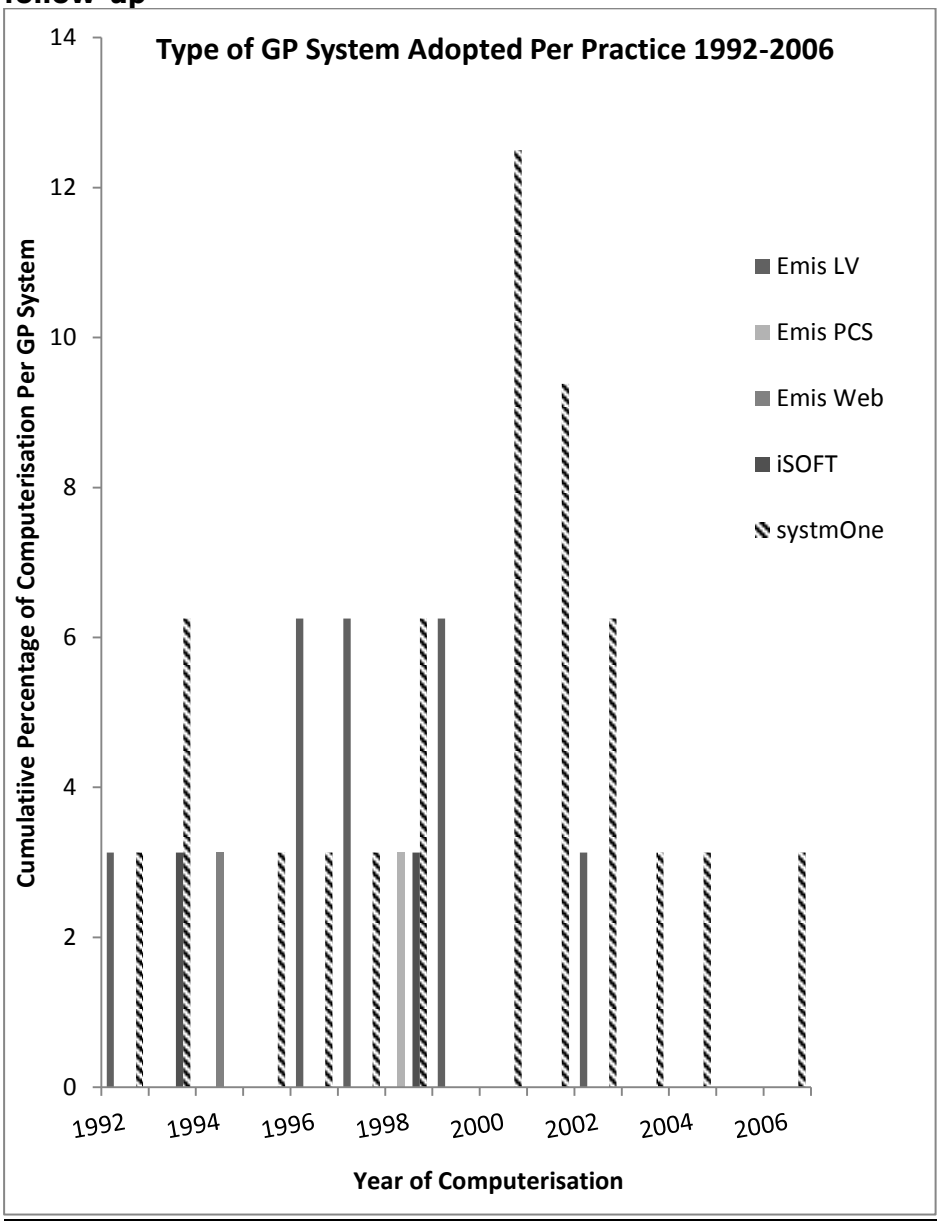
As the study is analysing data from EMRs a determination of a computerisation date is necessary so that any extracted data is compared during a time interval that can be reasonably sure that the respective EMRs were actually being used at the practice level. It is needed to be sure that analyses of extracted data when considering whether post computerisation has an effect occurs during a period where we can be reasonably sure that practice staff are recording using electronic means resulting in a positive level of record completeness. As a result for each practice in the fully-processed dataset, a computerisation date was determined. This was done by plotting the annual number of recordings per patient to determine the point at which the recording level reached 10% of the eventual plateau as shown in Figure 50 . The computerisation date was set to the 1st of January in the relevant year as shown in Figure 50. Using this computerisation date means it will be possible to use replicate analysis of EMRs allowing for post computerisation of the records; for example to assess if this had an impact on free text entry, prescriptions and consultations recorded.

Figure 50 determining the computerisation date of January 1st in relevant year by using asymptotic peak recording level as a proxy



It is also possible to determine a detailed view of the progress of computerisation for each practice encountered by using the same method above and breaking this down further according to practice and year and including the GP system found. Figure 51 shows that by 2004 most practices had already introduced and were using a GP system.

Figure 51 GP System Adopted by Practice for those encountered during data collection from consented patients during the 2007 follow-up



The three largest contributing GP software systems are systmOne, iSOFT and Emis LV. The next section presents the results from the EMR extraction and looks at free text, prescriptions and consultation measures and compares these across all GP software systems.

5.4 Results

It was possible to extract sex for all participants except for those participants from Emis LV (n=76 participants) and Emis Web (n=4 participants) meaning 80 participants' sex data value was not present in their records. It is possible that this data item was recorded elsewhere in the GP system but it was not present in the EMR extract that is used to transfer to another GP as if they are moving practice. It is possible to impute this value from manual validation of the record but this was not considered at this point in time as only data from EMRs was used. The results from analysing EMRs are given next.

5.4.1 Post computerisation adjustments for freetext, prescriptions and consultations

The median scores were re-calculated per patient allowing for post computerisation of GP records for freetext, prescriptions and consultations. This was to enable a comparison and examine any further possible changes to the median scores allowing for age, sex, IMD, social class once computerisation of EMRs occurred.

5.4.2 Free text

The median amount of free text for the systmOne was 35 (IQR 19-67) per consultation followed by iSOFT which had 11 (IQR 2-36). Emis PCS was the next highest figure with a median of 15 (IQR 8-27), Emis LV followed next with a median of 6 (IQR 1-15) and lastly Emis Web had a median of 4 (0-16) shown in table 10. A further statistical test on free text showed that there is a significant difference in the amount of free text per consultation between the different GP software systems ($p<0.001$). The same result was gained when free text was considered post computerisation. The median results post computerisation were virtually the same even when age, sex, socio-economic status were allowed for.

For the two largest contributing GP systems of systmOne and iSOFT it was found that older patients tended to have more free text per consultation. This was not seen compared to the same age ranges for Emis LV. Results show that the median calculated for each GP system did not vary by any large degree when broken down by age, sex, socio-economic status and social class. This remained the case post computerisation of records. The median free text values for age, sex, deprivation and social class generally remained consistent with the median for the respective GP system overall (i.e. when these variables were not considered separately) shown in Table 9 and Table 10.

Table 9 free text words per person per consultation by vendor, age, IMD and social class

	systmOne		*iSOFT*		*Emis LV*		*Emis PCS*		*Emis Web*			
	median (IQR=%)	(N) population	median (IQR=%)	(N) population	median (IQR=%)	(N) population	median (IQR=%)	(N) population	median (IQR=%)	(N) population	p-value	p-value for extra variables
Vendor	35 (19-67)	482	11 (2-36)	144	6 (1-15)	76	15 (8-27)	10	4 (0-16)	4	0.001	
Age:												0.001
30 < 49	28 (15-53)	39	8 (3-29)	6	7 (1-17)	6	0	0	4 (0-18)	1		
50 < 65	31 (17-60)	179	10 (1-35)	33	6 (1-15)	33	15 (8-28)	5	3 (0-18)	1		
66 > 75	36 (19-68)	151	11 (3-34)	16	6 (0-15)	16	15 (8-26)	4	13 (0-34)	1		
76 +	38 (20-75)	113	13 (3-42)	21	6 (1-14)	21	14 (7-26)	1	4 (0-9)	1		
Sex												0.001
Male	35 (19-68)	221	14 (4-43)	35	0		14 (8-28)	4	0	0		
Female	34 (18-66)	261	10 (2-32)	40	0		16 (8-27)	6	0	0		
missing					6 (1-15)	76			4 (0-16)	4		
Quintiles of IMD												0.001
Least deprived 1	36 (20-70)	106	10 (2-36)	22	7 (1-17)	22	0	0	4 (0-13)	3		
2	34 (18-63)	101	12 (3-36)	24	7 (2-14)	10	15 (8-23)	1	0	0		
3	35 (19-67)	85	11 (2-36)	47	5 (1-11)	12	17 (8-30)	1	0	0		
4	35 (19-66)	92	11 (3-32)	23	4 (1-10)	15	15 (8-27)	4	13 (0-34)	1		
Most deprived 5	34 (18-68)	93	12 (2-38)	27	8 (2-17)	15	16 (7-28)	4	0	0		
missing	25 (14-46)	5	11 (5-32)	1	15 (4-23)	2	0	0	0	0		
Social Class												0.001
Highest I	34 (18-63)	41	8 (1-30)	10	8 (2-18)	11	15 (8-23)	1	3 (0-18)	1		
II	34 (19-67)	108	13 (2-42)	43	6 (1-16)	21	17 (7-29)	1	0	0		
III	36 (19-70)	75	11 (3-32)	27	6 (1-14)	14	16 (8-28)	4	4 (0-18)	1		
IV	36 (18-70)	64	14 (2-49)	14	3 (0-8)	4	15 (9-28)	1	0	0		
Lowest V	34 (18-60)	6	9 (2-37)	1	0	0	0	0	0	0		
missing	34 (18-66)	188	10 (2-33)	49	6 (1-15)	26	15 (8-27)	3	4 (0-12)	2		

Table 10 free text words per person per consultation by vendor, age, IMD and social class post computerisation

	SystmOne		*iSOFT*		*Emis LV*		*Emis PCS*		*Emis Web*			
	median (IQR=%)	(N) population	median (IQR=%)	(N) population	median (IQR=%)	(N) population	median (IQR=%)	(N) population	median (IQR=%)	(N) population	p-value	p-value for extra variables
Vendor	37 (19-71)	482	14 (4-43)	144	8 (3-17)	76	16 (8-29)	10	4 (0-16)	4	0.001	
Age:												
30 < 49	31 (16-57)	39	10 (4-38)	9	9 (3-19)	6	0	0	8 (0-19)	1		
50 < 65	33 (18-65)	179	14 (3-47)	52	8 (3-17)	33	16 (8-29)	5	3 (0-18)	1		
66 > 75	37 (20-71)	151	13 (4-38)	51	8 (2-17)	16	15 (8-28)	4	13 (0-34)	1		
76 +	41 (21-80)	113	16 (6-49)	32	8 (3-17)	21	15 (7-30)	1	4 (0-9)	1		
Sex												0.001
Male	37 (20-72)	221	16 (6-49)	35			14 (8-28)	4				
Female	36 (19-71)	261	12 (3-39)	40			17 (8-29)	6				
missing					8 (3-17)	76			4 (0-16)	4		
Quintiles of IMD												0.001
Least deprived 1	38 (21-76)	106	13 (3-43)	22	9 (3-19)	22	0	0	4 (0-14)	3		
2	35 (18-66)	101	14 (5-43)	24	9 (4-17)	24	17 (8-24)	1	0	0		
3	38 (20-71)	85	15 (4-46)	47	6 (2-14)	47	18 (8-31)	1	0	0		
4	37 (19-69)	92	12 (4-37)	23	6 (2-15)	23	15 (8-28)	4	13 (0-34)	1		
Most deprived 5	36 (19-72)	93	14 (4-46)	27	10 (3-18)	27	16 (8-29)	4	0	0		
missing	25 (15-46)	5	12 (5-34)	1	16 (6-24)	1	0	0	0	0		
Social Class												0.001
Highest I	36 (19-69)	41	12 (2-38)	10	9 (3-19)	10	14 (8-24)	1	3 (0-18)	1		
II	36 (19-71)	108	15 (5-48)	43	8 (3-18)	43	17 (7-33)	1	0	0		
III	38 (20-75)	75	13 (5-36)	27	7 (2-16)	27	17 (8-30)	4	8 (0-19)	1		
IV	38 (19-74)	64	18 (5-57)	14	4 (0-8)	14	15 (9-28)	1	0	0		
Lowest V	36 (20-62)	6	15 (2-50)	1	0	1	0	0	0	0		
missing	37 (19-70)	188	13 (4-42)	49	8 (3-18)	49	16 (8-28)	3	4 (0-13)	2		

5.4.3 Prescriptions

The median number of prescriptions per person for systmOne was 113 (IQR 34-298), the next largest GP system was iSOFT 75 (IQR 32-183) and the smallest was Emis Web's median of 12 (IQR 9-30) shown in Table 11. The result showed that there is a statistically significant difference in the number of prescriptions across the GP software systems ($p<0.001$). The median prescriptions per year for patients post computerisation of GP records are shown in Table 12. These are virtually the same as Table 11 with results not showing significance when IMD quintiles and social class were allowed for both pre and post computerisation analysis (Table 11 and Table 12). Age showed statistically significant difference between the different GP software systems ($p<0.001$). Results indicate that for all GP systems older participants have a significantly higher number of prescriptions. This remained consistent when post computerised records were used per person per year. Females had a much higher number of prescriptions as a median value compared to men in all years and was the same post computerisation of records. There was no statistical significance in prescriptions when IMD and social class were allowed for as shown in Table 11. However for both socio-economic variables of IMD and social class showed that for the GP systems of systmOne and iSOFT the most deprived quintiles had a much higher number of prescriptions compared to the median prescription value for the two GP systems overall i.e. not broken down by no other factors. The results indicate an almost J-shaped relationship with the well middle social class having the least number of prescriptions. This was also true for Emis LV software system where a slightly less deprived quintiles of 2 and 3 showed lower median than the least and most deprived groups. The results for IMD and social class were again consistent in the same manner once computerisation of records was factored in and prescriptions re-calculated per year per person shown in Table 12 .

Table 11 total prescriptions per person per practice by vendor, age, IMD and social class: 1991-Collection Date

	systmOne		*iSOFT*		*Emis LV*		*Emis PCS*		*Emis Web*			
	median (IQR=%)	(N) population	median (IQR=%)	(N) population	median (IQR=%)	(N) population	median (IQR=%)	(N) population	median (IQR=%)	(N) population	p-value	p-value for extra variables
Vendor	113 (34-298)	482	75 (32-183)	144	26 (13 – 47)	76	44 (17-71)	10	12 (9 – 30)	4	0.001	
Age:												0.001
30-49	34 (11-119)	39	49 (28-146)	9	13.5 (7-26)	6	0	0	12 (12-12)	1		
50-65	59(18-153)	179	42 (24-104)	52	15 (7-41)	33	20 (17-71)	5	11 (11-11)	1		
66-75	129 (52-288)	151	103 (48-183)	51	27 (15-37)	16	44 (21-51)	4	6 (6-6)	1		
76+	343 (153-557)	113	140 (69-223)	32	43 (33-92)	21	78 (78-78)	1	47 (47-47)	1		
Sex												0.001
Male	100 (25-300)	221	50 (17-110)	35			56 (21-75)	4	0	0		
Female	129 (48-296)	261	134 (49-215)	40			34 (17-55)	6	0	0		
missing					26 (13 – 47)	76			12 (9 – 30)	4		
Quintiles of IMD												0.028
Least deprived 1	124 (48-336)	106	67 (37-149)	22	20 (13-41)	22	0	0	12 (11-47)	3		
2	109 (39-264)	101	76 (33-164)	24	30 (13-46)	10	47 (47-47)	1	0	0		
3	56(24-289)	85	56 (25-120)	47	26 (12.5-49)	12	55(55-55)	1	0	0		
4	116(32-291)	92	49 (23-146)	23	48 (12-60)	15	56 (29-92)	4	6 (6-6)	1		
Most deprived 5	153(52-359)	93	149 (84-205)	27	23 (10-42)	15	19(9-49)	4	0	0		
missing	25(13-43)	5	521 (521-521)	1	16(6-26)	2	0	0	0	0		
Social Class												0.237
Highest I	105 (34-282)	41	33 (19-77)	10	40 (13-49)	11	20 (20-20)	1	11 (11-11)	1		
II	94 (31-267)	108	56 (31-191)	43	25 (13-49)	21	17 (17-17)	1	0	0		
III	136 (51-357)	75	66 (21-188)	27	33 (19-41)	14	32 (9-63)	4	12 (12-12)	1		
IV	106 (34-382)	64	98 (30-134)	14	19 (8-35)	4	71 (71-71)	1	0	0		
Lowest V	227 (73-514)	6	434 (434-434)	1	0	0	0	0	0	0		
missing	119 (32-286)	188	104 (40-170)	49	21 (10-55)	26	55 (41-113)	3	27 (6-47)	2		

Table 12 prescriptions per person per year per practice by vendor, age, IMD and social class: post computerisation

	systmOne		*iSOFT*		*Emis LV*		*Emis PCS*		*Emis Web*			
	median (IQR=%) prescription per person year	(N) population	median (IQR=%) prescription per person year	(N) population	median (IQR=%) prescription per person year	(N) population	median (IQR=%) prescription per person year	(N) population	median (IQR=%) prescription per person year	(N) population	p-value	p-value for extra vraiables
Vendor	8.78 (2.35 - 22.89)	482	3.90 (1.72 - 9.40)	144	1.93 (0.90 - 3.17)	76	3.05 (1.21 - 4.75)	10	0.73 (0.49 - 1.68)	4	0.001	
Age:												0.001
30 < 49	2.08 (0.76 - 7.36)	39	1.86 (1.42 - 3.68)	9	1.13 (0.73 - 1.53)	6	0	0	0.87 (0.87 - 0.87)	1		
50 < 65	3.78 (1.27 - 12.87)	179	2.13 (1.08 - 5.80)	52	1.40 (0.48 - 2.50)	33	1.42 (1.21 - 4.75)	5	0.58 (0.58 - 0.58)	1		
66 > 75	10.41 (3.73 - 23.15)	151	5.52 (2.53 - 11.04)	51	1.76 (1.16 - 2.70)	16	3.05 (1.45 - 3.37)	4	0.39 (0.39 - 0.39)	1		
76 +	24.64 (11.88 - 44.85)	113	9.13 (3.75 - 15.31)	32	3.40 (2.13 - 6.09)	21	5.46 (5.46 - 5.46)	1	2.49 (2.49 - 2.49)	1		
Sex												0.001
Male	7.86 (1.58 - 23.45)	221	3.17 (1.27 - 6.65)	35		0	3.79 (1.45 - 5.11)	4		0		
Female	10.12 (2.96 - 22.27)	261	6.59 (1.94 - 10.76)	40		0	2.34 (1.21 - 3.48)	6		0		
missing					1.93 (0.90 - 3.17)	76			0.73 (0.49 - 1.68)	4		
Quintiles of IMD												0.028
Least deprived 1	7.87 (2.44 - 23.45)	106	3.75 (1.95 - 9.33)	22	1.51 (0.87 - 3.13)	22	0	0	0.87 (0.58 - 2.49)	3		
2	7.91 (2.59 - 21.61)	101	4.00 (1.79 - 8.20)	24	2.07 (1.58 - 2.73)	9	3.26 (3.26 - 3.26)	1	0	0		
3	4.98 (1.88 - 19.92)	85	2.58 (1.15 - 7.03)	47	1.79 (0.72 - 3.18)	12	3.48 (3.48 - 3.48)	1	0	0		
4	10.54 (2.56 - 21.76)	92	2.53 (1.08 - 8.96)	23	2.50 (1.36 - 3.53)	13	3.79 (2.02 - 6.35)	4	0.39 (0.39 - 0.39)	1		
Most deprived 5	16.01 (4.02 - 30.97)	93	8.75 (4.23 - 14.71)	27	1.58 (0.90 - 2.13)	13	1.31 (0.64 - 3.44)	4	0	0		
missing	1.80 (0.88 - 7.36)	5	26.62 (26.62 - 26.62)	1	1.29 (0.48 - 2.09)	7	0	0	0	0		
Social Class												0.237
Highest I	7.48 (2.44 - 24.31)	41	1.47 (0.74 - 3.53)	10	2.61 (1.13 - 3.60)	10	1.42 (1.42 - 1.42)	1	0.58 (0.58 - 0.58)	1		
II	7.43 (1.84 - 20.20)	108	3.51 (1.58 - 10.33)	43	2.06 (0.92 - 3.29)	21	1.21 (1.21 - 1.21)	1	0	0		
III	9.66 (3.45 - 28.49)	75	2.95 (1.08 - 11.83)	27	2.49 (1.40 - 2.89)	12	2.23 (0.64 - 4.36)	4	0.87 (0.87 - 0.87)	1		
IV	10.88 (2.62 - 23.65)	64	5.83 (1.42 - 8.07)	14	1.24 (0.59 - 1.79)	4	4.75 (4.75 - 4.75)	1	0	0		
Lowest V	16.14 (5.34 - 29.57)	6	28.25 (28.25 - 28.25)	1	0	0	0	0	0	0		
missing	8.66 (2.17 - 21.93)	189	5.52 (2.48 - 9.18)	49	1.55 (0.66 - 2.83)	29	3.48 (2.84 - 7.94)	3	1.44 (0.39 - 2.49)	2		

5.4.4 Consultations

The results showed that there is a statistically significant difference in the number of consultations between the different GP software systems ($p<0.001$). Results were also statistically significant in consultations across GP systems according to both age ($p<0.001$) and sex ($p<0.001$). SystmOne had a median for consultations per person of 169 (IQR 92-274). iSOFT's median score was 92 (IQR 58-147), the next largest contributing GP system of Emis LV had a median of 73 (IQR 41-105). These results are shown in Table 13. The median scores for consultations per person per year post computerisation of GP records are shown in Table 14. Emis Web and Emis PCS are excluded from the analysis as there is not enough data to comment on results but are displayed in the table for completion purposes. Across all GP systems older patients had higher number of consultations. This remained the case when medians for consultations were re-calculated per person per year post computerisation. The two largest contributing GP systems to the cohort are systmOne and iSOFT. For both of these GP systems women had a higher number of consultations compared to men. This was consistent post computerisation as well. No data on sex for Emis LV was available. Consultations were then considered using IMD deprivation and social class. The results were not statistically significant when IMD quintiles and social class were separately allowed for in both pre and post computerisation analyses. There were some notable but not significant results within the GP systems themselves where systmOne's least deprived quintile had a median of 204 (IQR 123-339) and its most deprived quintile's median was 200 (IQR 115-295). These were very similar to each other but both above the median for that GP system itself overall when not broken down by any other variable. This again suggests that consultations vary by IMD in a horizontal "J" shaped pattern; the same as prescriptions for systmOne. For iSOFT the least deprived

quintile had a lower median and the most deprived quintile had a higher median of consultations compared to that calculated for iSOFT overall when no other variables considered. In Emis LV the least and most deprived quintiles of participants were similar to the median calculated for the system overall and the middle deprivation quintiles from 2 to 5 were not uniform indicating a double slope in a "W" shaped fashion between least and most deprived quintiles. For the other socio-economic variable of social class systmOne participants in the highest social class had a median of 163 (IQR 103-264) which was similar to the value for systmOne overall (when no other factors were considered). For the lowest social class systmOne's median was 209 (IQR 174-329) and iSOFT's median for the same quintile was 210 (IQR 210-210). SystmOne showed a clear distinction by social class where lower social class showed a much higher number of consultations compared to highest social class and for the GP system overall where no other variables were considered. This distinction was also apparent in the results for the iSOFT GP software system. However there was a less clear distinction in Emis LV where it appears participants with lower social class have less consultations. Emis LV's highest social class had a median of 66 (IQR 35-99) and the lower social class of 4 had a median of 48 (IQR 19-96). A median on the lowest social class was not available. Both highest and lowest social class were lower than the median found for Emis LV overall when no other variables were considered. This remained consistent post computerisation of records and when age, sex, IMD and social class were considered and is shown in Table 14.

Table 13 total consultations per person per practice by vendor, age, IMD and social class: 1991-Collection Date

	systmOne		*iSOFT*		*Emis LV*		*Emis PCS*		*Emis Web*			
	median (IQR=%)	(N) population	median (IQR=%)	(N) population	median (IQR=%)	(N) population	median (IQR=%)	(N) population	median (IQR=%)	(N) population	p-value	p-value for extra variables
Vendor	169 (92-274)	483	92 (58-147)	144	73 (41-105)	76	49 (24-74)	10	9 (7-13)	4	0.001	
Age:												0.001
30-49	89 (58-210)	39	83 (57-122)	9	49 (22-90)	6	0	0	9 (9-9)	1		
50-65	116 (67-209)	180	76 (45-109)	52	51 (35-96)	33	45 (24-100)	5	9 (9-9)	1		
66-75	173 (116-274)	151	93 (59-156)	51	76 (50-87)	16	44 (22-63)	4	4 (4-4)	1		
76+	270 (182-405)	113	117 (81-185)	32	98 (70-139)	21	65 (65-65)	1	17 (17-17)	1		
Sex												0.001
Male	149 (73-264)	221	66 (41-107)	72	0	0	70 (37-87)	4	0	0		
Female	181 (110-276)	262	119 (77-157)	72	0	0	40 (24-52)	6	0	0		
Quintiles of IMD												0.006
Least deprived 1	204 (123-339)	106	82 (44-133)	22	79 (36-98)	22	95 (54-153)	24	9 (9-9)	3		
2	161 (97-236)	101	0	0	55 (40-88)	10	0	0	0	0		
3	119 (68-264)	85	88 (58-130)	47	69 (44-119)	12	52 (52-52)	1	0	0		
4	155 (86-259)	93	75 (48-154)	23	91 (50-127)	15	87 (49-138)	4	4 (4-4)	1		
Most deprived 5	200 (115-295)	93	120 (87-173)	27	70 (29-117)	15	35 (16-55)	4	0	0		
missing	89 (78-94)	5	373 (373-373)	1	73 (66-8)	2	0	0	0	0		
By Social Class												0.281
Highest I	163 (103-264)	41	74 (40-100)	10	66 (35-99)	11	45 (45-45)	1	9 (9-9)	1		
II	155 (78-258)	108	81 (50-153)	43	65 (36-112)	21	24 (24-24)	1	0	0		
III	187 (97-295)	75	95 (61-161)	27	81 (50-96)	14	30 (16-50)	4	9 (9-9)	1		
IV	166 (88-307)	64	80 (62-126)	14	48 (19-96)	4	100 (100-100	1	0	0		
Lowest V	209 (174-329)	6	210 (210-210)	1	0	0	0	0	0	0		
missing	170 (96-266)	189	114 (59-158)	49	78 (42-106)	26	74 (52-176)	3	11 (4-17)	2		

Table 14 consultations per person per year per practice by vendor, age, IMD and social class: post computerisation

	systmOne		*iSOFT*		*Emis LV*		*Emis PCS*		*Emis Web*		p-value	p-value on extra variables
	median (IQR=%) consultations per person year	(N) population	median (IQR=%) consultations per person year	(N) population	median (IQR=%) consultations per person year	(N) population	median (IQR=%) consultations per person year	(N) population	median (IQR=%) consultations per person year	(N) population		
Vendor	12.75 (6.25 - 21.08)	483	4.79 (2.88 - 7.29)	144	4.34 (2.33 - 6.41)	76	2.91 (1.49 - 5.04)	10	0.46 (0.35 - 0.66)	4	0.001	
Age:												0.001
30 < 49	6.70 (3.30 - 14.93)	39	3.51 (2.85 - 4.06)	9	3.48 (1.80 - 6.83)	6	0	0	0.43 (0.43 - 0.43)	1		
50 < 65	8.92 (4.24 - 15.11)	179	3.72 (2.04 - 5.47)	52	2.84 (1.82 - 6.94)	33	2.48 (1.49 - 6.38)	5	0.48 (0.48 - 0.48)	1		
66 > 75	13.06 (7.97 - 21.32)	151	4.90 (3.32 - 7.49)	51	4.40 (3.71 - 5.35)	16	2.66 (1.24 - 4.18)	4	0.26 (0.26 - 0.26)	1		
76 +	21.34 (13.93 - 28.82)	113	6.90 (4.58 - 10.18)	32	5.63 (3.49 - 7.76)	21	3.76 (3.76 - 3.76)	1	0.85 (0.85 - 0.85)	1		
Sex												0.001
Male	11.19 (5.10 - 21.34)	221	4.00 (2.10 - 5.92)	35			4.40 (2.13 - 5.71)	4				
Female	12.97 (7.11 - 20.31)	261	5.58 (3.59 - 9.30)	40			2.23 (1.49 - 3.33)	6				
missing					4.34 (2.33 - 6.41)	76			0.46 (0.35 - 0.66)	4		
Quintiles of IMD												0.006
Least deprived 1	12.67 (7.30 - 21.71)	106	4.27 (3.22 - 6.33)	22	5.43 (2.33 - 6.96)	22	0	0	0.48 (0.43 - 0.85)	3		
2	11.54 (6.17 - 17.27)	101	5.01 (2.19 - 7.41)	24	3.25 (2.00 - 4.80)	10	1.99 (1.99 - 1.99)	1	0	0		
3	9.77 (4.81 - 20.25)	85	4.22 (2.90 - 6.96)	47	3.97 (1.92 - 6.62)	12	3.33 (3.33 - 3.33)	1	0	0		
4	13.98 (6.31 - 19.18)	93	3.64 (2.43 - 8.44)	23	4.45 (3.42 - 6.39)	13	5.71 (3.26 - 8.62)	4	0.26 (0.26 - 0.26)	1		
Most deprived 5	16.17 (9.41 - 23.54)	93	6.75 (5.38 - 9.32)	27	3.78 (1.80 - 6.28)	14	1.88 (0.89 - 3.12)	4	0	0		
missing	6.32 (5.10 - 16.08)	5	18.72 (18.72 - 18.72)	1	5.26 (4.58 - 5.95)	2	0	0	0	0		
Social Class												0.281
Highest I	10.89 (6.96 - 22.84)	41	3.12 (1.58 - 5.11)	10	5.57 (2.65 - 7.31)	10	2.48 (2.48 - 2.48)	1	0.48 (0.48 - 0.48)	1		
II	12.13 (5.92 - 19.54)	108	4.45 (2.85 - 7.38)	43	4.25 (1.82 - 6.83)	21	1.49 (1.49 - 1.49)	1	0	0		
III	14.14 (7.24 - 22.51)	75	4.53 (2.95 - 8.70)	27	3.97 (2.26 - 6.41)	14	1.63 (0.89 - 2.87)	4	0.43 (0.43 - 0.43)	1		
IV	13.93 (5.87 - 20.32)	64	5.56 (3.32 - 7.12)	14	2.82 (1.31 - 5.06)	4	6.38 (6.38 - 6.38)	1	0	0		
Lowest V	13.76 (12.07 - 19.18)	6	12.12 (12.12 - 12.12)	1	0	0	0	0	0	0		
missing	12.75 (5.57 - 19.44)	188	5.16 (3.23 - 7.86)	49	4.40 (2.51 - 6.45)	27	5.04 (3.33 - 10.85)	3	0.55 (0.26 - 0.85)	2		

5.5 Discussion

5.5.1 Principal Findings

- The type of GP system being used played a significant role in determining the amount of free text being used at the practice. This remained significant with age, sex, IMD, social class and including post computerisation of records ($p<0.001$).
- The type of GP system being used played a significant role in determining the amount of prescriptions ($p<0.001$). This remained significant with age ($p<0.001$) and sex ($p<0.001$). This remained consistent post computerisation across all the GP systems.
- The type of GP system being used played a significant role in determining the amount of consultations received by the participant ($p<0.001$). This remained significant with age ($p<0.001$) and sex ($p<0.001$). This remained consistent post computerisation across all the GP systems.
- Older participants had a significantly greater amount of free text
- Females had a significantly higher number of prescriptions and consultations compared to males

<u>Other Observed Differences between GP Systems</u>

- Greater deprivation did not significantly increase propensity to receive prescriptions or consultations at practices
- Lower social classes received a greater number of prescriptions and consultations in the two largest contributing GP systems of systmOne and iSOFT
- Prescriptions and consultations vary with females having a slightly higher number of prescriptions and consultations.
- Older age groups have a marked increase in prescriptions and consultations compared to younger participants.

5.6 Comparison with other studies

Analysis of GP records has shown variation across the different GP software systems found which at their core are all professing to do the same thing for doctors and patients in being a medium for recording. Yet results show that differences in their respective records is having a statistical significance between around free text, prescriptions and consultations. There were other non-significant but notable differences across free text, prescriptions and consultations between the GP software systems by age, sex deprivation and social class. However it would be unwise to attribute all differences to the design of GP systems themselves. The role the practice level human factor both in terms of patients and GPs and their staff needs to be considered within this context. The differences found and any links between them are discussed next in the context of existing studies.

5.6.1 Free text

It is important to determine whether free text differs across GP systems as it's a potential source of hidden medical data which may or may not be recorded elsewhere in the EMR. If we know the differences in free text between systems we can make allowances once data is extracted. We can speculate what could be stored in free text but there is no way to be certain but unstructured text suggests a human language narrative which may refer to some coded items. The variation found in our results comparing free text between the GP systems could be attributed to dependencies at the user interface level how it is used by staff. For example it can be expected that end users would enter much more of their own text if the user interface of a GP system encourages greater keyboard input. Some interfaces can encourage 'velocity coding' where the most frequently used codes are prioritised at the top and then picked proportionally higher than other codes (68). Such codes are likely to be those that apply to the largest number of patients by indicating a broad category of disease, rather than

specific subtypes, necessating the recording of further important details in the free text. The problem is compounded by the Read coding system being complex allowing a disease to be coded in many different ways, through a diagnostic code, or indirectly via history, symptoms or procedures. This can lead to wide variations in the way in which general practitioners code clinical problems (71), and the ratio between the different types of recording (and, consequently, the additional content that must be noted in the free text) is likely to be influenced by the relative ease of coding in each manner in different software packages. Regardless of the system in use, free text has an important role to play as one recent study found that potential disease cases may be missed or wrongly dated when coded data alone are used to identify patients was diagnostic suspicions are frequently confined to free text (149). This study found that older people have more free text found in their EMRs. Existing studies have also used free text in highly specific diseases and outcomes. For example a study that looked into how much information on ovarian cancer diagnosis is hidden in free text found that in most cases free text confirmed the coded diagnosis but that GPs did not code a definite diagnosis on the date that ovarian cancer was confirmed (146). This matters because clinical workflows and the corresponding use GP systems could explain why older patients more likely to suffer from things like cancer diagnosis and secondary care meaning hospital letters and automated updates between primary and secondary care are all contributing to older patients having more free text because they are more likely to be sicker from progressive illnesses requiring secondary hospital treatment and monitoring. Existing literature of free text has generally looked at making use of free text to examine the extent of how much uncoded and useful information is contained therein. One study has stated that the difficulty in using free text is that it contains a wide range of terminology, complex language structures, context-specific abbreviations and acronyms (150). Another study found that free text is valuable in obtaining more detailed information on cases but should not be used without manual case validation

(151). One study on free text use concluded that it can also help the task of increased cohort identification using a large clinical dataset so being able to reduce big data into a manageable size also requires careful manual review (152). This shows that from an extraction point of view free text is not the best method in which to record EMR data as its unstructured and ambiguous and it cannot be turned it into epidemiological measurements easily. It is also resource heavy to de-identify free text to share and disseminate. The study has found that free text is present in all the collected EMRs in different places and we can conclude that due to this pervasive nature it will be part of EMRs for the foreseeable future. Existing studies assessing different outcomes have successfully used it to draw conclusions on cancer recording and cohort identification and this study has shown that it is possible to identify this free text in various GP systems and use it to make meaningful comparisons across those systems and this can add to the picture of patient care because free text can be analysed according to age, sex, deprivation and social class with the added perspective of different GP systems on their own as well.

5.6.2 Prescriptions

The study found differences in prescription rate by GP system, and when broken down by age, sex, deprivation and social class. The receipt of drugs should primarily be dictated by patient needs, so it is unlikely that a GP software system would cause patients to be prescribed fewer or more drugs in reality. One potential explanation is the geographic clustering of patients by software vendor; the larger practices (with greatest numbers of study participants) tended to use SystmOne and may have been located in areas where patients experience poorer health and require more medication, however if this were the major factor it would be expected that the median values would be similar once the population is stratified by demographic factors. The fact that differences persist once this is done, suggest that alternative factors also play a role.

It may be that practices using certain systems coincidentally differ in the willingness of GPs to medicate patients and that the observed differences are a consequence of the relatively small number of practices in this study failing to "average out" such differences. Alternatively, it is possible that the EMR could hold the same prescription information differently between the GP systems. For example, each system can be configured to auto-fill details such as dosage and quantity of medication with default values. The British National Formulary (BNF) doesn't always advise on optimum duration of treatment so it is possible GP software companies see gaps in areas for which they can automate tasks for GPs and unwittingly influence the amount of medication prescribed, and therefore the number of additional prescriptions required when the first course of treatment does not completely resolve an illness. Similarly, some packages may encourage different patterns of issuance (e.g. monthly vs. quarterly) of long-term repeat prescriptions, plausibly leading to very large differences in the overall numbers over the time period of the EMR. Overcoming such shortcomings in prescription data requires further work to the processed data set to look at actual quantities of pills prescribed but this would require further data cleansing. In that situation there would also be a need to account for all the different pack/pill sizes when working this out and it would infeasible to do for all the drugs and would instead focus on a subset. Differences in prescription rates by age, sex, deprivation and social class could be better explained if there was specificity of prescriptions in terms of the type of problem they had aimed to resolve. This study does not seek to look at specific prescriptions solving diagnoses but similar studies that looked at using EMR-prescription data focused on anti-depressant prescribing and found that practice data does not provide enough information on allergic reactions to allow discernment between true allergies and drug side effects (139). This is relevant because to go beyond crude measures of prescribing rates the nature of problem and its corresponding prescription should be looked at together to examine demographic and clinical

trends in a more detailed analysis to see if there are then differences in demography and age for example. This would be a truer and fairer examination of EMR prescription data. The same view was supported by another study that also looked anti-depressants in electronic prescription records and was able to use EMR data to get prevalence and incidence in a cohort (117). That study found that one of the main problems associated with automated prescription data from EMRs is that it is subject to misclassification. Results from this study and the existing literature shows there are various influences on the prescription data that could account for the differences encountered in prescription rates across the GP systems.

5.6.3 Consultations

Similar trends to prescriptions were found in the consultation rates across the GP systems for age, sex, deprivation and social class. Again it is unlikely that different computer systems would affect the number of GP visits in real life, unless there are auto-prompts for screening follow-ups for example. This tends to support the suggestion that these differences may be partly driven by systematic differences between the practices such as the overall health of their populations, or the clinical preferences of the practitioners.

Another possibility is that the use of uniquely dated recordings is a poor measure of actual consultations; additional unique dates may appear due to automated additions to records. For example automatic lab links where results from lab tests come in and are recorded but these are not real consultations but might still be chronologically needed on screen. Automatic generation of repeats and any other automated entry triggered by admin roles such as hospital feedback or discharge letters received by email/secure link could influence consultation rates. All the GP software packages do offer these facilities but not all practices may use them or implement in the same manner. It can give rise to the situation where a small number of practices with some specific software settings and a small number of patients mean this could have a large and

overstated effect on these results. Furthermore, even when all such facilities are in use, packages may differ in terms of how these automatic additions are dated – information may automatically be added as it is received (triggering a new date in the record) or be held until the patient's record is next opened (e.g. during a consultation), effectively coalescing multiple recordings into a single date.

This possibility is supported in a recent study that looked at EMR data in diabetes patients and found that extracted data did not match expected prevalence and suggested that the various data extraction techniques they tested were missing large numbers of relevant patients (153). They concluded that the data extraction models they considered were ineffective for measuring the quality of diabetes care but were acceptable for crude extract numbers. They suggested that GP software systems should be subject to independent accreditation and review and monitoring for critical incidents associated with their use and cited another study (154) supporting greater accountability and transparency from software companies that make GP systems. The NHS has attempted to do some of this by setting standards for those practices ordering via their funded procurement process GP Systems of Choice (GPSoC) and covered earlier in chapter 3. However GPSoC standards were not created with epidemiological analyses in mind so that EMR limitations of misclassification and varied clinical workflows effecting prescribing and recording are still valid arguments.

5.6.4 Differences in median values for prescriptions and consultations

The median and IQRs for prescriptions and consultation illustrate that a similar trend is seen in age groups where older patients have higher number of prescriptions. It is seen in systmOne and iSOFT but less obvious in Emis editions. IMD and Social class indicate a general trend where the least deprived and higher social class tend to have a greater number of prescriptions. There are similar patterns seen in Emis packages. The median IQRs show that the

patterns observed for prescriptions are broadly similar for consultations across age, sex and deprivation. This shows that prescriptions and consultations are related enough for the key patterns identified by age, sex and IMD to be similar for prescriptions and consultations. Again it is unlikely that different computer systems would affect the number of GP visits or prescriptions with the most probable reasons for differences being how the various GP systems store similar data in their EMRs and how their respective software reacts to clinical workflows in its own environment and/or local customisations by staff.

These issues could be examined in more detail by comparing patterns of prescribing of particular medications, or patterns of consultation for specific diseases among the different software systems. It would then be possible to calculate and compare the actual amounts of drug received over a given period of time, or to examine recordings relating to a condition in detail to differentiate between those that are suggestive of an actual visit vs. those that could have been generated automatically. This was not attempted in the present study as the relatively small numbers of patients whose records were recorded using each system (other than SystmOne) makes this impractical – for a relatively common disease or medication with period prevalence of 5%, even the ISOFT population would only be expected to produce 7 or 8 patients.

5.6.5 The relationship between free text, prescriptions and consultations in the different GP systems

This study has found that free text, prescriptions and consultations are related to one another because they follow similar patterns when compared across the different GP systems. A possible explanation for the relationship could lie in the clinical workflow at practice level and the GP systems influencing such workflow and prescribing and recording habits and local customisations of GP systems. For example an explanation of the results could be that there is a link between a high number of consultations and higher amounts of free text because both

could be caused by automatic entries produced by the specific GP systems themselves and one system may encourage such text input as well. Another explanation could be that if certain problems are considered minor because they are not incentivised they may be more likely to be recorded in free text as there will be no remuneration for accurate coding and those practice have a high burden and GPs limited by time for each patient. Something similar could very much account for differences in prescription vs consultation rates where it is possible that a consultation could take place but the potential drug prescription could be brought cheaper over the counter. Certainly more affluent patients not in receipt of any free prescriptions may opt for this so making them appear to be healthier as they have lower number of prescriptions recorded. Existing studies have supported the notion that better interface design of EMRs at the user level means more accurate EMRs and improved quality of care (155,156). This would also directly affect free text, prescriptions and consultation data as well. For example one study found that several user interface changes to a GP software package reduced the incidence of prescription errors (157). The same study indicated that the most common reason for prescription discrepancies was due to differences between structured drop down lists and free text components. This means that if the human process of how prescribing and recording occurs in EMRs is closely scrutinised by GP software developers and assessed for quality during software development then it can result in safer care for patients and accurate rate of prescriptions and consultations data. It is also possible that better recorded prescriptions and accurately coded problems during consultations might have an effect of reduced need to use free text in EMRs as well. To prove these beyond doubt would require working with the GP software developers themselves and establishing a group of test practices and compare with practices using an unchanged user interface and then see the effect of interface changes over time and on free text, prescriptions and consultations. This is unrealistic. However a reasonable recommendation can be made that GP

software systems should have a nationally set of standards under schemes like GPSoC for example could certainly extend its requirements to the detail of the human computer interface of GP software systems being properly tested with real end users. There is no evidence from looking at software developers websites that they actually do test with real end users. This is challenging with practices that cannot spare staff for user testing due to the burden of caring for patients. It is possible that some GP software companies do test with real GP practice staff. It is difficult to pinpoint highly specific reasons for differences in free text, prescriptions and consultations across the GP systems but it is probably a combination of factors and circumstances such as staff training, GP habits, default settings on GP software systems, poorly understood user documentation, local customisations and practice culture and patient care influences that contribute for such disparities.

5.6.6 Implications for using EHRS from GP systems for research

- It is possible to create your own database specific to a cohort and extract all the data from their EMRs into a working database
- The extraction software extracted all data from EMRs in a highly accurate and adequate way and this new software can potentially be used in other studies with minimal to no changes and reused freely in EMR analysis.
- Free text can be defined in a certain measurement (e.g. number of words per consultation) for a cohort. There is no universal or gold standard of measurement of free text as it seems to be dependent on each study's outcomes and specific classification of what is and what is not free text.
- Variation in free text, prescriptions consultations are linked with each other and vary by age and sex and socio-economic status. The type of GP systems employed seems to play a significant role in how this occurs by influencing the structure of their respective EMRs.

- Privacy and costs of de-identification in free text will have to be considered for the extraction software to be used in a safe manner.
- Extraction software can directly be used in an existing cohort following up EMR data at practices to get a detailed view of their care at a fixed point in time using free text, prescription and consultation rates as comparison measures.
- Pattern matching can take place on drug names, dates of consultation and use the demographics to achieve view on deprivation levels as well
- Specific measures can be extracted for quality control from EMRs as the extraction software was able to extract all Read codes found in the EMRs tested.

5.7 Strengths and Limitations

Strengths

Manual checking of the extracted software demonstrated that the software extraction program accurately extracted all the data from collected EMRs. There were similar patterns for free text, consultations and prescriptions by age, sex and social class that suggest and increased validity of the data extraction process across the different GP systems. At each practice the method used was consistent based on same scenario of patients moving to another practice.

Limitations

In some of the GP systems there is limited data for some patients. For example if they moved practice and not all previous free text was entered as accurately as problems and prescriptions. Another limitation is that regular expressions that make up the extraction software will not work very well if the GP software vendors start making changes to the structure of their electronic records as the regular expressions depend on detecting patterns to know what to extract. There is no way to verify how the different practices use free text even though we are using it as a comparison indicator. For example high free text could

mean better recording of information or less well recorded prescriptions if a practice uses free text to record a minor prescription. For example there could be bias that is difficult to detect could have be introduced in using free text if at a practice level episodes that not incentivised are not recorded in the same manner but are found in free text. Another potential problem is that the length of repeated prescriptions can be different in different practices for certain chronic conditions. For example this could make it appear that certain groups more likely to have chronic conditions would appear to have different levels of prescriptions compared to consultations.

5.8 Summary

There is variation across different GP software systems all professing to do the same thing for doctors and patients yet showing differences in how they structure their respective records. There were differences noted across free text, prescriptions and consultations between the GP systems. However it would be unfair to attribute all differences to the design of GP systems themselves without working with the GP software companies and real life GP staff. The role the practice level human factors both in terms of patients and practitioners and their staff needs to be considered within this context. The key findings from this extraction and comparison are:

- Free text varies considerably between different GP systems.
- Within age, sex, deprivation and social class there are not major differences within the type of GP system and EMRs are from different practices.
- Prescriptions vary within age, sex, deprivation and social class. With Females having a slightly higher number of prescriptions and older age groups a marked increase in prescriptions compared to younger participants.

- Consultations fit a similar pattern to prescriptions across vendors and age, sex, deprivation and social class.

There can be many reasons for these differences both from clinical work flow and GP recording habits. The study has shown that EMRs can be analysed and compared from different GP systems.

6 Comparison of Extracted Medical Recordings of the Gedling Database to Cohort Questionnaire Follow-ups 1991, 2000 and 2007

Aim:

Compare diagnosis of four different outcomes in Gedling study EMR data to data collected in follow-up surveys in 1991, 2000, and 2007.

Objectives:

- Compare self-reported doctor diagnosis of asthma, hay fever, and eczema compared to data recorded in the EMR.
- Using data from the Forced Expiratory Volume over 1 second determined diagnosis of COPD and compare to the data recorded in the electronic health records.

6.1 Introduction

This chapter will examine the recordings in the collected GP records and compare these to self-reported and observed recordings collected during the Gedling study follow-ups of 1991, 2000 and 2007. It further extends the research on extraction from GP records by analysing the extracted database files and comparing it to known previous recordings not related to the GP record. It will allow an examination of the consistency of GP recording and forms an exemplary first use of the Gedling database in a real world setting and to demonstrate the feasibility of using EMR-derived database in actual cohort that has been followed since 1991.

6.2 Background

It is important that the EMRs the results are based on are as complete and accurate as possible. If the EMRs show discrepancies then this has a major bearing on whether extracted data can ever be used in analysis and form the basis of sound judgement. For example one study found that better electronic

coding of diabetes patients was associated with improved care and that some patients were incorrectly coded and therefore wrongly excluded from the appropriate disease register and missed out on care surveillance needed (158). This chapter focuses on the area of completeness as a comparison to known cohort data as this will help build the solid foundations for long term and effective use of EMRs for ongoing clinical research. The inconsistencies in completeness in EMRs exist and are a central factor to handle. For example one study found that sicker patients tended to have more complete records and healthier patients tended to have records that were less complete (159). That study concluded a word of caution in wishing to reuse such EMR-derived databases as blind sampling of complete records within an EMR database may skew the sampled population towards sicker patients. For example another study concluded that EMR completeness would contribute greatly to disease estimation because epidemiologic studies which use medical records to verify self-reported drug exposure may in fact underestimate the number of omitted prescriptions and so more complete medical records would better reflect true patient histories (160). The same study warned against going to the other extreme and forcing data entry can introduce biases that are difficult to detect or correct. It means that there is potential for EMR data to be reused effectively if it is an accurate account of patients' health. For example another study indicated that the outcome of using EMR data to examine community-acquired pneumonia showed that the apparently healthiest patients died at a higher rate than sicker patients because the results were bias due patients who came to the emergency department and died quickly did not have many symptoms entered into their EMR (161). Better complete GP records have an effect on better patient care in terms of accurate EMRs, properly maintained diseases registers, that people who need treatment don't miss out and prescriptions that have fewer errors. All this will contribute greater validity of results and effect of conclusions derived.

6.3 Data Collected from Previous Follow-ups

Table 15 gives an overview of the data collected from previous surveys and what is available from looking at the electronic medical records collected for Gedling cohort during 2011-2012. The previous Gedling surveys asked patients specific questions on their health and whether they had certain conditions. This was in addition to other measurements taken at the time as well. The following problems were specifically asked whether they felt it had been recorded in their notes and patients self-reported for the following outcomes:

- Asthma
- Wheeze
- Eczema and or rash
- Hay fever
- Smoking status
- Age

Chronic obstructive pulmonary disease was not self-reported although lung function tests were measured and recorded. Height recordings were measured for the patients during the 1991 and 2000 follow-up.

6.3.1 Review of Gedling Survey Measurements 1991-2007

Table 15 Data available from previous Gedling Follow-ups 1991, 2000 and 2007 and data available in Gedling cohort electronic medical records collected

Outcome	Gedling Survey Year			Electronic medical record collection 2011-2012
	1991	2000	2007	available or coded in electronic medical record collected
Asthma diagnosis Confirmed by Doctor	Patient self-reported	Patient self-reported	Electronic Health Record examined	☑
Eczema or Rash	Patient self-reported	Patient self-reported	Electronic Health Record examined	☑
Hay fever	Patient self-reported	Patient self-reported	Patient self-reported	☑
Chronic obstructive pulmonary disease	Can be calculated from spirometry results	Can be calculated from spirometry results	Not measured	☑
Skin prick tests to known skin allergens	Measured at follow-up	Measured at follow-up	Not measured	Not available
Forced Expiratory Volume over 1 second (FEV_1)	Measured at follow-up	Measured at follow-up	Not measured	☑
Airway reactivity to methacholine measured by a drop in FEV_1	Measured at follow-up	Not measured	Not measured	Not available
Food frequency questionnaire	Measured at follow-up	Measured at follow-up	Measured at follow-up	Not available
Smoking status	Patient self-reported	Patient self-reported	Patient self-reported	☑
Age	Patient self-reported	Patient self-reported	Not measured	☑
Sex	Patient self-reported	Patient self-reported	Not measured	☑
Height	Measured at survey	Measured at survey	Not measured	☑
Wheezing	Patient self-reported	Patient self-reported	Patient self-reported	☑

6.4 Asthma and Chronic Obstructive Pulmonary Disease

We do not have the statistical power to make any significant findings looking at asthma and COPD achievement in the Gedling database. This chapter looks at these outcomes from the perspective of comparing between self-reported data collected during previous Gedling cohort follow-ups what is contained in the Gedling database formed from their EMRs collected at GP practices

6.5 Rationale

Studies validating existing large databases have shown poor agreement with patient self-report for some outcomes (162,163). The Gedling data sets produced from the data extraction allow us to partly assess GP record quality because we can compare them to self-report surveys and measurements taken during previous follow-ups. The next section presents the findings from analysing the Gedling EMR-derived database and comparing to independently existing known data for the Gedling cohort follow up data collected from 1991, 2000 and 2007. This chapter will look at the following self-reported outcomes from previous Gedling cohort follows:

- Asthma

COPD will also be considered as an outcome but was not self-reported:

- Chronic obstructive pulmonary disease cases were identified by case finding from lung measurements collected in 1991 and 2000.

6.6 Methods

6.6.1 Dataset & Study Population

This is defined as those patients for whom we were able to collect GP records. A detailed breakdown of this population is provided in chapter 3. The reason for this definition is that this the population from which the data sets in the Gedling database is derived.

6.6.2 Defining the outcome code lists to examine Gedling cohort GP recordings

6.6.2.1 Extracting Clinical Codes for Read version 2 and Clinical Terms Version 3

By comparing with previous research it is possible to use validated Read code lists where Read version 2 (and other clinical codes like CTV3) were used to define asthma (164) and COPD (165) in general practice, clinical code-lists were created for both of these outcomes. These code lists for asthma and COPD are included in appendix 8.11

6.6.3 Statistical Methods and Analyses

The set of validated clinical codes (Read 2 and CTV3) for asthma and COPD were used to identify cases of earliest diagnosis in the Gedling database for each EMR collected. During analyses only unique clinical codes were used to identify the relevant diagnoses for the different problems. Any duplicate clinical codes from the code lists given in appendix 8.11 were dropped prior to identifying earliest diagnoses for the outcomes. The survey dates used in computerised records were:

- 1st January 1991
- 1st January 2000
- 1st January 2007

The Read code field from the Gedling database was merged with the clinical code list for each problem being considered of asthma and COPD. Once the earliest recording of problems was identified a comparison was then made by looking at the equivalent recordings in the self-reported health status data collected during previous Gedling cohort follow-ups of 1991, 2000 and 2007. The basic algorithm was that the earliest Read or CTV3 clinical code of each outcome (asthma and COPD) recording before each of the three survey dates (as above for 1991, 2000 and 2007) was compared to the self-reported outcome in each survey.

6.6.3.1 COPD Case Finding

COPD diagnosis was done by "COPD case finding through lung function measurements collected during 1991 and 2000 follow-ups. COPD diagnosis is done by looking at the percentage of predicted forced expiratory volume in 1 second (FEV_1) as recommended in the National Institute for Health and Clinical Excellence (NICE) 2010 guidelines for the management of COPD in adults (166). According to their guidelines FEV_1 >80% of predicted with a Read code for COPD was classed as mild and then the following categories applied to FEV_1:

- 50-80% moderate,
- 30-50% severe and
- 30% very severe

This study found COPD cases based on measured lung function data and GOLD criteria 2 or above using 1991 and 2000 cohorts. Grading of severity of airflow limitation in COPD (based on post-bronchodilator FEV_1). In patients with FEV1/FVC (Forced Vital Capacity) 0.70: GOLD 2: moderate 50% < FEV1 80% predicted as a COPD case found. To ensure clearer derivation of COPD cases in addition to obstructive spirometry smoking history of 10 pack years and shortness of breath/cough were used to identify COPD cases from the questionnaire data and measurements collected in 1991 and 2000 follow-ups. There was also an extra analysis which did not include a specific point in time but looked at whether there was a presence or absence of a recording of an outcome diagnoses in GP records at any time in the past up to point of EMR data collection at practices. For COPD detection in EMRs a validated code list of COPD codes were used to identify COPD in GP records from a previous study (165).

6.6.3.2 Sensitivity Analysis using Drug Recordings for Asthma

As a further test of asthma recordings in the Gedling database and self-reported asthma in 1991, 2000 and 2007 follow-ups asthma drug codes were used to test asthma recordings in EMRs when compared with self-reported asthma in 1991,

2000 and 2007 follow-ups. The reason for using asthma drug recordings as well as clinical codes is due to it being a severe health problem when considered alongside other self-reported problems of wheeze, eczema or rash and hay fever. COPD was identified through case finding (not self-reported) and there is overlap between drugs for both asthma and COPD so only asthma recording was further tested using drug prescriptions in EMRs collected and using this with clinical codes during time intervals between 1991, 2000 and 2007. The asthma drug code list used for searching the Gedling database for was from a case-control study which looked at inhaled corticosteroids and the heightened risk of pneumonia in asthma cases (167). This study was chosen as it was peer reviewed and is a recent study by year and also looked using primary care data to examine asthma cases. The asthma drug code list used is provided in appendix 8.11. In the initial analysis the earliest recording of asthma using clinical codes is considered. The asthma drugs list is used instead of clinical codes to make a comparison between the two ways of assessing recording. The earliest asthma drug recording is used instead of earliest asthma diagnoses clinical codes. This is then used to compare to self-reported asthma in same way as clinical codes. A further test of sensitivity is applied by combining drug prescriptions and clinical codes of asthma diagnoses in the following manner for the three survey dates:

- An asthma Read/CTV3 code before Jan 1991 and an asthma drug record before 1991 compared to self-report in 1991 follow-up.
- An asthma Read/CTV3 code before Jan 2000 and an asthma drug between years 1991 < 2000 compared to self-report in 2000 follow-up.
- An asthma Read/CTV3 code before Jan 2007 and asthma drug recording between years 2000 < 2007 compared to self-report in 2007 follow-up.

It was useful to use asthma drugs in this way as it provided a sensitivity analysis alongside the clinical codes for Read version 2 and CTV3. A tightening to the definition of asthma recording provides more than one view on recordings over the same time periods.

6.7 Results

Patients' responses from answers given to previous cohort follow-up questionnaires was then cross tabulated with the diagnoses extracted from their EMRs. For COPD this was measured by case finding from 1991 and 2000 compared with EMR coded diagnoses during the same years. The results are presented below for the outcomes of asthma and COPD.

6.7.1 Self-reported asthma compared to asthma diagnoses in electronic medical records 1991, 2000 and 2007

Results in Table 16 show that around 24% of those who reported to have asthma in the 1991 Gedling cohort had the diagnosis recorded in their EMR prior to 1991. In the 2000 follow-up this figure had risen to 44% of patients reporting asthma had a recording of an asthma diagnosis before the year 200. In the 2007 follow-up about 75% of patients who reported having asthma had a diagnosis for the condition before the year 2007.

Table 16 diagnoses of asthma in electronic medical records compared to Doctor diagnosed asthma at successive Gedling cohort surveys 1991, 2000 and 2007

		Doctor-reported asthma at 1991 survey	
Asthma recorded in electronic medical record before 1991		Yes (%)	No (%)
	Yes (%)	10 (23.81)	4 (0.60)
	No (%)	32 (76.19)	661 (99.40)
	At Any Time		
	Yes (%)	25 (59.52)	54 (8.12)
	No (%)	17 (40.48)	611 (91.88)

		Doctor-reported asthma at 2000 survey	
Asthma recorded in electronic medical record before 2000		Yes (%)	No (%)
	Yes (%)	20 (44.44)	5 (1.23)
	No (%)	25 (55.56)	402 (98.77)
	At Any Time		
	Yes (%)	29 (64.44)	18 (4.42)
	No (%)	16 (35.56)	389 (95.58)

		Doctor-reported asthma at 2007 survey	
Asthma recorded in electronic medical record before 2007		Yes (%)	No (%)
	Yes (%)	62 (74.70)	11 (1.79)
	No (%)	21 (25.30)	605 (98.21)
	At Any Time		
	Yes (%)	63 (75.90)	17 (2.76)
	No (%)	20 (24.10)	599 (97.24)

6.7.2 Asthma drugs recorded in electronic medical records compared to doctor-reported asthma at successive Gedling cohort surveys 1991, 2000 and 2007

Results in Table 17 show that in the 1991 cohort 14% of those reporting doctor-diagnosed asthma had a recording of an asthma drug in their EMRs before 1991. In the 2000 follow-up 53% of patients who had reported doctor-diagnosed asthma contained an asthma drug recording in their EMR before 2000. In the 2007 follow-up this figure was 72% of those reporting to have doctor-diagnosed asthma containing an asthma drug recording in their EMRs before the year 2007.

6.7.3 Diagnoses of asthma with a recording of an asthma drug in electronic medical records compared to doctor-reported asthma at successive Gedling cohort surveys 1991, 2000 and 2007

Results in Table 18 show that in the 1991 cohort 5% of those reporting to have doctor-diagnosed asthma had both an asthma diagnosis and an asthma drug recording in their EMRs before 1991. This had risen by the 2000 follow-up, where 20% of patients who reported to have doctor-diagnosed asthma had both an asthma diagnosis and an asthma drug recording in their EMRs between the years 1991 and 2000. By the 2007 follow-up this figure was 51% of those reporting to have doctor-diagnosed asthma had a diagnosis for asthma and an asthma drug recording between 2000 and 2007.

Table 17 Asthma drug recordings in electronic medical records compared to doctor-reported asthma in questionnaire results 1991, 2000 and 2007

		Doctor-reported asthma at 1991 survey	
Asthma drugs prescribed in electronic medical record		Yes (%)	No (%)
Before 01/01/1991			
	Yes (%)	6 (14.29)	11 (1.35)
	No (%)	36 (85.71)	654 (98.35)
At Any Time			
	Yes (%)	25 (59.52)	39 (5.86)
	No (%)	17 (40.48)	626 (94.14)

		Doctor-reported asthma at 2000 survey	
Asthma drugs prescribed in electronic medical record		Yes (%)	No (%)
Before 01/01/2000			
	Yes (%)	24 (53.33)	11 (2.70)
	No (%)	21 (46.67)	396 (97.30)
At Any Time			
	Yes (%)	24 (53.33)	395 (97.05)
	No (%)	21 (46.67)	12 (2.95)

		Doctor-reported asthma at 2007 survey	
Asthma drugs prescribed in electronic medical record		Yes (%)	No (%)
Before 01/01/2007			
	Yes (%)	56 (71.79)	10 (1.69)
	No (%)	22 (28.21)	580 (98.31)
At Any Time			
	Yes (%)	54 (69.23)	9 (1.53)
	No (%)	24 (30.77)	581 (98.47)

Table 18 - Diagnoses of asthma with a recording of an asthma drug in electronic patient records versus doctor-reported asthma at successive Gedling cohort surveys 1991, 2000 and 2007

	Doctor-reported asthma at 1991 survey	
Asthma recorded in electronic medical record before survey and with an asthma drug prior to 1991	Yes (%)	No (%)
Yes (%)	2 (4.76)	3 (0.45)
No (%)	40 (95.24)	662 (99.55)
At Any Time		
Yes (%)	25 (59.52)	40 (6.02)
No (%)	17 (40.48)	625 (93.98)

	Doctor-reported asthma at 2000 survey	
Asthma recorded in electronic medical record before 2000 and with an asthma drug between years 1991 < 2000	Yes (%)	No (%)
Yes (%)	9 (20.00)	2 (0.49)
No (%)	36 (80.00)	405 (99.51)
At Any Time		
Yes (%)	25 (55.56)	12 (2.95)
No (%)	20 (44.44)	395 (97.05)

	Doctor-reported asthma at 2007 survey	
Asthma recorded in electronic medical record before 2007 and with an asthma drug between years 2000 < 2007	Yes (%)	No (%)
Yes (%)	40 (51.28)	6 (1.02)
No (%)	38 (48.72)	584 (98.98)
At Any Time		
Yes (%)	55 (70.51)	9 (1.53)
No (%)	23 (29.49)	581 (98.47)

6.7.4 Cases found for COPD from measurements collected in 1991 and 2000 compared to COPD diagnoses found in electronic medical records

Results in Table 19 show that COPD cases found from the 1991 cohort around 96% did not have COPD diagnoses before 1991. In the 2000 follow-up all the COPD cases found did not have any recorded diagnoses before the year 2000. When any recorded diagnosis of COPD is considered in 1991 38% of the cases found in the 1991 cohort have a recording of COPD. In 2000 of the 28% COPD cases found had a COPD diagnoses at any time in their EMRs.

Table 19 Diagnoses of Chronic Obstructive Pulmonary Disease (COPD) in electronic patient records compared to case finding of COPD using Forced Expiratory Volume 1 (FEV1) recorded at successive Gedling cohort surveys 1991 and 2000

		COPD cases found in 1991 survey	
COPD recorded in electronic medical record before 1991		Yes (%)	No (%)
	Yes (%)	1 (4.17)	0 (0.00)
	No (%)	23 (95.83)	683 (100.00)
	At Any Time		
	Yes (%)	9 (37.50)	25 (3.66)
	No (%)	15 (62.50)	658 (96.34)

		COPD cases found in 2000 survey	
COPD recorded in electronic medical record before 2000		Yes (%)	No (%)
	Yes (%)	0 (0.00)	1 (0.24)
	No (%)	29 (100.00)	422 (99.76)
	At Any Time		
	Yes (%)	8 (27.59)	15 (3.55)
	No (%)	21 (72.41)	408 (96.45)

6.8 Discussion

6.8.1 Principal Findings

- There is a large increase in EMR asthma recordings compared to doctor-reported asthma between 1991, 2000 and 2007. There was improvement between all the time periods.
- Using earliest asthma drug recording in place of clinical asthma codes resulted in a substantial reduction in percentage of asthma recording compared to doctor-reported asthma. This suggests using asthma clinical codes are better for looking at earliest asthma recordings in EMRs.
- Using asthma clinical codes combined with asthma drug codes to examine drug recording between follow-up intervals showed a greater drop in asthma recording compared to doctor-reported asthma.
- Tightening the definition of asthma beyond earliest clinical codes compared to doctor-reported does suggest that in the early computerisation of GP records asthma drugs could have been prescribed on paper and this diagnosis of asthma was not recorded until a later exacerbation.
- COPD recording in EMRs compared to case finding in 1991 and 2000 showed that a very high percentage of cases were not recorded and this did not improve between the time periods. COPD recording in EMRs compared to case finding was found to be poor.

6.8.2 Strengths and Limitations

A major strength of this data are that by using the complete GP records until date of collection we have a reasonably accurate account of the GP recordings over time since initial computerisation which should accurately reflect the participants' medical history. This means that the EMR data are not subject to participant recall bias. However it is possible that some seemingly minor conditions could be self-treated with over the counter medication and this data would never get recorded in their EMRs leading to underestimation of the

recordings for less serious problems. A limitation could be that we do not have access to EMRs from previous computer systems as the current GP software system may not have always been at the practice. It is impossible to know whether all the data was properly transferred during system migration or whether manual re-entry onto the new system was needed and whether this meant that only serious disorders were given priority or misclassification of existing problems occurred during any such process. One limitation could be that the diagnosis may only have been recorded at an early time point prior to computerisation and a second recording may not have occurred until a serious exacerbation of the problem such as asthma. Meaning it would be possible for participants to correctly recall but this not be coded onto the computerised for a long interval. We have mitigated this to a certain degree by using "at any time" to maximise comparison of recall to the whole period of EMRs rather than just at specific intervals. One strength of using a variety of clinical codes is that it softens the problem of various GP systems over time and different practices using different definitions of clinical codes with different released versions. There is anecdotal evidence from visiting a practice where we informed that they had not updated their code lists for latest version of QOF from their IT provider. Using Read version 2, CTV3 and QOF-codes maximises the identification of cases mitigating slightly any minor misclassification using a comprehensive list of clinical codes that cover the UK standard coding schemes. However a limitation is that obsolete codes such as the Oxmis system were not used in the analyses as they were not part of the TRUD release pack from HSCIC and no mappings were found based on any previous studies to allow accurate validation of equivalent Oxmis codes. Another limitation when looking at both asthma and COPD is that there is some diagnostic uncertainty or overlap between asthma and COPD. For example patients who had records of both were not considered separately. The reason for this is that we are looking at recording of the condition over the follow-up time intervals and it is possible to have both at the

same time rather than one or the other. Another approach to making comparisons between the different GP systems could have been to compare a specific drug for asthma as an example and examine the prevalence of recording in THIN and make a comparison with the Gedling database recording for the same drug or disease. However there are limitations that exist for such an approach. For example selecting a drug that would have a reasonable prevalence in both data sets. Not everyone could be prescribed the same type of drug even across the different age spans making comparisons difficult between THIN and the Gedling database. Another limitation affecting comparisons would be that the Gedling database is a selection population in terms of a long standing cohort in a specific geography of Nottinghamshire with no Vision GP software contributing data to the study making the population too specific for direct comparisons to THIN. One limitation of the COPD code list used from the 2014 COPD study (165) was that the original study used version 2 Read codes mapped from their original CPRD codes. However version 3 CTV3 codes were not used in that specific study so had to be mapped from the version 2 Read codes using the TRUD dictionary (141). Another limitation to COPD case finding was that data was not available for shortness of breath in 1991.

6.8.3 Findings in Context of Previous Research

6.8.3.1 Asthma Clinical and Drug Recordings

However this study also found that once asthma prescriptions were included for an asthma case then the percentage of asthma recording compared to self-reported asthma did reduce on all three years from 1991 to 2007. This shows that prescriptions can play an important role in determining the actual prevalence or burden-related factors such as health costs of a certain disease when using EMR data in addition to Read and CTV3 coded diagnoses. For example one study found that dispensed asthmatic medication data would be a useful proxy for the presence of current asthma when disease register data are

not available (168). This result could be because patients that had asthma diagnosed in the past but did not need any drugs for it in the interval periods of the survey. It is possible that the drop in recorded asthma once drug recordings were combined could mean that some patients may actually be classed as asthma cases but may fail to self-report this as asthma due to it being stable and not actively using asthma drugs and wrongly assuming they do not have asthma. It is also possible they were given a drug recording of asthma but not officially classed as asthma suffers due to a stable condition and did not return to see their GP until a later exacerbation. The findings of a drop in asthma recording compared to self-reported when asthma drugs are used in between the time intervals combined with the presence of a clinical code during the same time period could be because once clinical code are used for diagnoses, any repeat visits centre around mainly prescribing drugs rather than any further specific clinical coding codes for the asthma consultation itself. It suggests that emphasis might be shifting to focusing on recording the treatment rather than including any further clinical asthma codes. For example one study interviewed patients and compared their allergy self-reporting to EMRs and found the percentage of agreement between interviews and charting for reaction type was around 50% concluding that better methods are needed to properly record allergies to ensure patient safety and care (169). There is also evidence to suggest the method used in terms of pick lists on GP software interfaces contributes considerably to variation of allergy recordings even leading to inaccurate recording where patients are denied treatments they are mistakenly recorded to be allergic to (170). However there is evidence to show that wheeze has been a source of false-positive results when looking for asthma in EMRs indicating that it does have some relationship with the diagnosis process for more serious respiratory disorders (171).

6.8.3.2 Chronic Obstructive Pulmonary Disease (COPD) and EMR recordings

A large number of participants in 2000 were found to have COPD but this was not recorded in their EMRs and the results were quite different from those found in 1991 which showed better concurrence between case finding and EMR recordings. This means that some cohort members in 2000 and 1991 would be otherwise be classed as COPD cases during those times but were missing out on monitoring and treatment. This could be because the patient population was getting older in 2000 but early disease did not manifest enough symptoms to warrant consultation with their GP. There is no case finding data available for the 2007 cohort. However there is evidence to get a good idea of what has happened to COPD recording in EMRs after 2000. For example a 2008 study found that the incentivised COPD targets under QOF since 2004 has had an impact in the prevalence of recorded diagnoses of COPD in the UK between 2003 and 2005. That same study also found that there was an increase in the recording of spirometry data and prescriptions for combination inhalers suggesting a beneficial impact on the care of people with COPD. It is a serious condition compared to other outcomes analysed such as hay fever and eczema with gradual worsening over time. Evidence there is now better recording of COPD and COPD-related interventions to the extent COPD patients can be accurately identified and acute exacerbations of COPD can be found (165,172). Earlier EMR recording of COPD is good from the patient healthcare perspective as there are findings from a study to suggest diagnosis of COPD is strongly associated with a diagnosis of lung cancer, this association explained by smoking habit and strongly dependent on the timing of COPD diagnosis (173). This can mean better recording of COPD disease management in EMRs can help in surveillance for any associated problems such as cancer and cardiovascular disease.

6.8.4 Implications of Findings

There are two drivers that have been major influences on the study findings when using extracted EMR data and self-reported outcomes including COPD case finding. It is possible that the influences come from two distinct areas which could be argued are symbiotic with one another. A balance has to be struck between allowing for real environment use having an effect on EMRs such as pressures on staff and still making use of the potential of EMRs in clinical research. In this situation the EMR data may not be completely perfect but it is data which holds value and can still be used for effectively. This study agrees with previous research that concluded researchers should adopt validated, systematic methods of EMR data quality assessment (130). This study also supports a previous study's conclusion there are multiple layers of EMR completeness which show different results (174) and the choice of an EMR attribute(s) under study is a major study decision. This is a significant implication for using EMRs in secondary databases such as CPRD and THIN in terms of incomplete or missing fields in EMRs at the GP practice level. However there is a previous study looking at accurately identifying COPD patients in major health databases that has shown a high level of accuracy since the incentivised scheme began in 2004 (165). This means that coded diagnoses in EMRs are being done accurately for COPD so are valid and useful for research.

Implementation of the Quality and Outcomes Framework in 2004

6.8.4.1 Increased use of electronic medical records at practices

QOF's introduction incentivised certain areas of healthcare and this is illustrated in the results where asthma shows a rapid recording of its diagnoses compared to other less serious outcomes like hay fever. COPD recording in EMRs showed that there were disparities in recording between 1991 and 2000 and improved recording before 2000 was not uniformly high. This may show that including COPD in the QOF scheme launch was a sensible approach as many people were found to have COPD during case finding but this was not recorded in their EMR data at that time. This shows that QOF may be creating a ripple effect of

awareness for simple symptoms that may be related to much more serious problems being better recorded. This would be a positive for the patient but the downside is an eagerness to classify a patient into a QOF ruleset. For example some serious illnesses can have no symptoms except a rash and/or gentle cough such as the Human Immunodeficiency Virus (HIV). There is no doubt that QOF has had a major impact on asthma recording (175,176) and COPD recording (177) respectively. However more research is needed on how diagnoses that are considered benign are recorded and whether they need to be incentivised. For example this is supported by a previous study that stated non-incentivised activities and patients' concerns may receive less clinical attention (178). However that study concluded that the improved organizational ability to deliver effective chronic disease management suggested that at least some of the reported improvements in performance were due to increased investment of staff time and much increased clinical activity, rather than simply better recording of existing behaviours. There is evidence to support this from another high-burden respiratory problem where a recent study found that practices with higher patient-reported access had lower rates of emergency admissions for asthma (179). This is unfortunate because symptoms seen as mundane can be indicative of serious future illness or result in missed opportunities in surveillance. This is an understandable phenomenon as QOF uses data from the primary care level to work out financial rewards to different practices using different GP software systems. But QOF's limitation is that it doesn't show the whole picture of an EMR and things are missing or misclassified as it doesn't cover everything in an EMR. This can result in missing data in EMRs which has implications for clinical research using the EMRs. Nearly all practices use EMRs and this means it's a lot easier to access electronic recordings of illnesses. However there is a lot of variation between GP systems. It is possible that these differences have an effect on the specific of recordings of illnesses as well. Recordings in EMRs for asthma showed a dramatic improvement between 1991

and 2007. QOF reporting of asthma started in 2004 (180) and looking at asthma in particular the early years of EMR recording is poor and recording increases substantially by 2007. It is a probable influence in better asthma EMR recording. For example it could be that GPs were recording asthma using QOF codes for relatively severe cases or those most suitable for QOF interventions. There can be other non-QOF reasons for these findings, one possible reason could be that the patients were wrong and made mistakes in filling out the survey. It is unlikely that such mistakes would be made on such a marked scale to adversely affect the results. Possibly a further explanation is that the early recording could be held in any free text areas of EMRs and GPs got better at recording asthma accurately with better GP systems and training. It may also be possible that less serious problems are recorded in free text and not allocated a clinical code. It is possible that these differences have an effect on the specific of recordings of illnesses as well.

6.8.5 Summary

This chapter looked at taking the outcomes of asthma and COPD and assessed the level of recording in EMR data collected in previous Gedling study follow-ups. In addition to QOF diagnoses codes for asthma and COPD from 2004/05 to 2011/12 a wider set of asthma and COPD codes covering the clinical coding schemes of Read version 2 and Clinical Terms version 3 were used from previous studies used as a comparison when constructing the code list. For greater sensitivity asthma drugs were also used to identify asthma in EMRs. These were chosen due to previously collected cohort data with which to make a comparison between actual EMR recordings and self-reported health status for 1991, 2000 and 2007 follow-ups. Once comparisons were complete it was possible to assess how well EMR recordings matched with self-reported outcomes during the different time intervals. Some outcomes were better recorded than others and

there is also recent evidence to make an assessment about recent EMR recording patterns.

7 Conclusions and Recommendations

To reiterate the aims of the thesis were as follows:

- This study aimed to use electronic health data in two different research projects. The initial aim of this thesis is to become familiar with the structure and format of research databases by carrying out an exemplar research study using an existing large database of primary care medical records (The Health Improvement Network database).
- The second aim was to obtain the EMRs of the Gedling cohort participants and to convert them into a standardised format suitable for use in future research projects. Thus allowing researchers to build bespoke primary care EMR databases for their own cohorts of patients by visiting patients' practices and extracting data directly from their EMR installation.

The key findings in response to the thesis aims are summarised in section 8.1. Section 8.2 outlines the implications of these findings and Section 8.4 discusses recommendations on the basis of this work for further studies and section 8.5 provides a concluding statement on the study.

7.1 Summary of Main Findings

7.1.1 Extraction from Primary Care Electronic Medical Records Collected from the Gedling Cohort into the Gedling Database

The study then built on the understanding of primary care data by using THIN as a template with which to examine the same outcomes of asthma and COPD (in addition to others) on practice data collected from a Nottingham-based cohort. Several mechanisms were considered to use for the data extraction and finally regular expressions were found to possess the key advantage of being able to look for highly specialised patterns of data. These patterns were then designed and then iteratively tested on randomly selected EMRs until the respective EMRs showed that their data was extracted into data tables based on THIN. The main finding was that it was possible to use regular expressions in a successful manner in terms of time taken to process all EMRs from the range of GP

software systems encountered in a matter of several minutes. Also successful mechanism in terms data accuracy and iterative testing because if there any data missing from EMRs during testing the regular expression could be fine-tuned in real time and the data extract tested rapidly. This aided successful programming to ensure accurate regex extractions took place. To ensure greater reliability two independent reviewers examined the data extracted into files compared to an unprocessed electronic version of the EMRs. Another important finding was that there very important differences between the GP systems when it comes to EMR layout, several synonyms for the type of medical information and differences in coding versions such as some practices using only Read version 2 and some that used Clinical Terms version three (CTV3).
Once all EMR data was extracted successfully a final set of data files formed the newly made Gedling database.

7.1.2 Comparison of the Extracted Data in the Gedling Database across the different GP Software Systems

The Gedling database was then utilised to examine differences between EMR measurements across the range of GP systems encountered. The measures chosen because of their basic function in EMRs is in recording consultation, issuing prescriptions and recording free text. We needed to examine what happened to these measurements when different GP software systems were included in the analyses. The main finding was that the type of GP software does make a significant difference in the amount of free text and this remains a significant when age, sex, deprivation and social class are included in analyses. There was more free text in older people, women and deprived populations. Generally older patients tend to have more free text per consultation in systmOne and iSOFT software. Whereas the variation in free text by age is not seen as significantly in older patients among the Emis editions of LV, PCS and Web software. It also played a significant difference in recording of prescriptions and consultations per patient and this remained significant for age and sex.

Deprivation and social class were not found to have much of an influence on prescriptions and consultations when different GP software systems were compared. Older patients and females were found to have a higher number of prescriptions. Generally the least deprived group tended to have a greater number of prescriptions though not significantly. Consultations were found to fit a similar pattern to prescriptions where older patients and females and least deprived groups had a greater a number of consultations. The main finding that GP systems do have an influence on important EMR facets of free text, prescriptions and consultations is significant and important for studies making use of EMR data derived databases especially if EMRs are from a variety of different software companies. This also remained the case when a computerisation date indicating full use of EMRs for recording was used in analyses. This investigation tested the potential of using small scale health databases from a specific cohort and examining their recording by type of GP software system that produced the EMR.

7.1.3 Comparison of Extracted Medical Recordings of the Gedling Database to Cohort Questionnaire Follow-ups 1991-2007

The Gedling database was then used to assess outcomes by examining asthma and COPD. These outcomes were analysed by looking at the GP recordings (case finding for COPD) to self-reported or GP recorded outcomes. The main finding was that that asthma and COPD recordings were close to self-reported outcomes. Main finding demonstrates that EMRs are increasingly being used for recording outcomes accurately. This has improved dramatically from year 2000 onwards. This investigation tested the potential of using small scale health databases from a specific cohort and examining their recording at practice level comparted to self-reported outcomes over a period of time.

7.2 Main Implications of the Study

7.2.1 Extracting data from EMRs collected at GP practices

Similar studies using EMRs and regular expressions with manual review has shown to provide an accurate mechanism of data extraction (135,181,182). This study adds to that nascent body of evidence that it is a practical route to utilise the same methods and an appropriateness for using in other situations. There are other mechanisms for data extraction that will be covered in the Recommendations for further studies section of the current chapter.

7.2.2 The effect of GP software on EMR recordings at practices

EMRs produced by the different GP systems have a unique layout depending on which GP software system is in use at practices. As this study has shown that this has an effect on results when considering free text, prescriptions and consultations. This means that studies that use different EMRs could produce inaccurate possibly skewed results due to large numbers of EMRs from a certain GP system that would magnify a certain aspect. For example one GP system may encourage use of automated templates that result in only certain codes and text being used. This study supports the view that the choice of GP systems do make a difference to the recording that takes place on an EMR. Thus supporting the view of a recent study which concluded that under the UK's pay-for-performance scheme, differences in practice performance were associated with the choice of clinical computing system stating that inconsistencies across systems need to be understood and addressed and caution being used when generalising findings from a single computing system (155). This is very similar to this study's experience when addressing these differences in the design and accuracy testing phase of the Gedling database all the way to differences found when the database was ultimately used to examine free text, prescriptions, consultations and finally GP recordings to self-reported outcomes. This study supports the view that the type of GP software system uses makes a definite impact on the areas we have studied and there are significant variability in the nature of what EMR data looks like from the various practices depending on the GP system used.

7.3 Recommendations for further studies

7.3.1 Large Scale Data Extraction

This was a relatively small scale study with a specific cohort whom we knew things about in advance from previous questionnaires. Future research therefore needs to contend with limitations that are significant if attempting to do something similar on a larger scale or different circumstances. There is much more potential for future research in EMRs to show that when larger areas of EMRs are examined and compared missing data can be found. This is important since any future research can be aware of potential pitfalls occurring if other dimensions of EMRs are subject to intense data mining and any conclusions drawn on those results can then be checked for validity. For example one study indicated that the outcome of using EMR data to examine community-acquired pneumonia showed that the apparently healthiest patients died at a higher rate than sicker patients because the results were bias due patients who came to the emergency department and died quickly did not have many symptoms entered into their EMR (161). But again the balance has to be struck between allowing for real environment use and pressures on staff and making use of potential of EMRs for clinical research. The same study warned against going to the other extreme and forcing data entry can introduce biases that are difficult to detect or correct. A possible recommendation based on this study is that access to underlying structure of a GP software design even just an abstracted view from a GP software company of the design within a certain research licence at least could help when developing extraction software and enable better bridging of EMR differences. Such better collaboration is a view was also supported by a recent study that stated higher levels of confidence in quality measures reporting at the population level by establishing collaborative user groups that work with EMR vendors as partners and use technical assistance to build relationships and trust in EMR-generated reports (156). That same study also

recommended that software vendors should continue to explore modifications for improving data capture (at the front-end) via standardised data entry templates. This study's future recommendation also supports this view that it would help greatly towards better standardisation of EMR data meaning extracts are much easier to analyse. Although this should not come at the expense of patient care in any way by limiting options at the practice level's interface.

7.3.2 Exploring Free text in EMRs

Another key area for future research using EMRs is in the area of free text and how to make the best use out of it. This study showed that there is lots of free text present and that it varies by age, sex and deprivation and GP software vendor and can be effectively extracted. However this study did not take it a step further and do extractions on the free text itself. A future research possibility could be to compare and contrast with the data quality from the structured part of the EMRs to free text. The downside of using regular expressions is that they are not easily suited to the unstructured nature of free text unless what is being sought is specific and structured in the midst of free text. By keeping an open mind on what is being sought to start with will ensure something that might be of use but not considered is not missed and discounted as just 'noise' inside EMRs. For example one study looked at using a free text matching algorithm to extract diagnoses and causes of death from unstructured text in EMRs and concluded that it was valid and reliable in their context of extracting coded information from free text in GP records with good precision (183). However again looking for a specific and coded structure in free text is comparatively easier but future research needs to move towards focusing on the surrounding context when it comes to making use of free text. This will allow future research to attempt to develop a sort of taxonomy when classifying free text as a beginning to exploring the possibilities and allowing better communication between clinical researchers as well. One recent study that used

free text to estimate the incidence of rheumatoid arthritis in primary care concluded that most studies are based on coded information whilst ignoring information stored in free text (149). They stipulated that potential cases may be missed or wrongly dated when coded data alone are used to identify patients as diagnostic suspicions are frequently confined to text only. Therefore their key conclusion was that use of EMRs to create disease registers or assess quality of care will be misleading if free text information is not taken into account. They recommended that methods to facilitate the automated processing of free text need to be developed and implemented. Future research in techniques to handle free text can open the door for EMR-extracted data to be exhibit greater validity and completeness. For example one study indicated that this key area of completeness in EMRs is context driven and illustrated multiple definitions of completeness being used, different degrees of measured completeness for the same dataset, and concluded that the number of complete records in a typical clinical database may be far lower than the nominal total (174). Therefore a future area of research for free text could examine whether greater amount of free text is a reliable proxy for EMR completeness and so study inclusion. That way any results from free text and structured aspects of EMRs will better handle the constraint of incomplete EMRs.

7.3.3 Using EMR recordings to assess patient care

Another area of future research could also look at whether there any other aspects of EMRs that can also be used as useful proxies such as prescriptions themselves. This study observed the variation of prescription rates by age, sex and deprivation. Future research for example could look at the nature of these prescriptions to see a different perspective on what is happening within a cohort being followed-up over time. For example one study found that mother-reported use of anti-asthmatics during the previous year observed among 7-year old children was highly valid and so concluded that dispensed anti-asthmatics

would be a useful proxy for the presence of current asthma when disease data are not available (168). Future research can focus on the area of completeness; as this will help build the solid foundations for long term and effective use of EMRs for ongoing clinical research. The discrepancies in completeness in EMRs exist and are a central factor to handle. For example one study found that sicker patients tended to have more complete records and healthier patients tended to have records that were less complete (159). That study concluded a word of caution in wishing to reuse such EMR-derived databases as blind sampling of complete records within an EMR database may skew the sampled population towards sicker patients. Future research can evaluate these types of confounding and contribute to our understanding of effects caused by EMR completeness. For example another study concluded that completeness would contribute greatly to disease estimation because epidemiologic studies which use medical records to verify self-reported drug exposure may in fact underestimate verification or underestimate the number of omitted prescriptions and so more complete medical records would permit the verification of true patient histories as well as the identification of false patient histories (160).

7.3.4 Encountering other primary care systems not covered

This study looked at extracting from GP software vendors encountered as the practices were visited. However, the Vision practice system was not encountered at all and so future research could look at extending the current software design so that it can handle patients from the Vision GP system. It could also look at other techniques to obtain data extraction that were not implemented. This study used regular expressions to identify and extract data but other methods do exist and have been applied. For example one study used a technique called Natural Language Processing (NLP) using General Architecture for Text Engineering (GATE) software to extract a variety of grammatical features, which were then used to create specific filters to maximize precision and recall on

instances of specific drug prescribing (123). Future research can look into using such Natural Language Processing as it can be used as an open source (cost-free) software licence. That study was also able to verify the validity of the data derived against a manually coded 'gold standard' to establish precision and recall measurements and they then went onto estimate the disease specific prevalence and prescription patterns. Therefore research could examine such techniques that are clinically useful and technically feasible. This study has shown one technique of using regular expressions to transform EMR data into knowledge. One recent study states that managing accessibility and privacy will be of key concern before such studies can be done on a larger scale (184). Throughout this study close attention was given to these matters when transporting, storing and analysing the data and high strength encryption was used in all instances. For clinical data in EMRs to be utilised properly on large scale assurances of security and privacy must be reliable for participants to be encouraged to take part. One area of future research could look into using regular expressions on a larger scale with GP records from different vendors and even different countries to examine if the results gained broadly match the patterns found in this study.

7.3.5 Combining clinical studies with GP data in future research

Another possible area of research was indicated in a study where it was suggested that it is possible to combine data from the EMR with other sources such as clinical data, environmental information, and gene sequencing data (184). This type of research would give a whole new perspective that would be able to continually change and adapt if linked with such data inputs. The ability to adapt to variation will be crucial feature for success as this study has found that variation between GP software systems was very much the norm and lots of time and resources were spent in bridging these divides. Future research could take large amounts of EMRs and create better linkages with existing health databases like THIN or CPRD, hospital notes, other disease registries and even

death records. According to one study this type of approach can potentially create an observational evidence base for clinical questions that would otherwise not be possible (185). The same study concluded with the example of a major drawback caused by this latter issue of 'generalisability' where it limits the application of conclusions derived from randomized trials performed on a narrow spectrum of participants to patients who exhibit very different characteristics. Overcoming such limits with data sets rich in EMR-derived and better linked measurements that are reliable can certainly be a possible area for further exploration. This could be achieved by linking existing health audit tools to make a much larger connected data set rich in clean data which is ready to use. For example Primis which was formed in 2000 and is business unit of the University of Nottingham aiming to improve patient care using general practice data of better quality and accessibility have produced various health audit tools for asthma, COPD, coeliac disease, atrial fibrillation and many others (186) could at some point could link their audited data to GP data sets such as the Gedling database to provide a new perspective on patients' health. It might be possible to interrogate the data in a new way or map the linked up data to market segmentation data such as Mosaic to see the characteristics of patients and their ailments to develop better targeted and timely interventions.

7.4 Concluding Remark on Fulfilment of Aims

This study has harnessed the power of extraction software to make it possible to research into EMRs that are a huge potential repository of medical data. These can be explored and utilised for testing established thinking gained from existing large health databases but thus far never in bespoke health databases from specific cohorts with previous questionnaire data as this study has done. The thesis aimed to use electronic health data in two different research projects. This was ultimately fulfilled in more than two areas. Subsequently using THIN as a template for the Gedling Database. Once the Gedling database was built it

was used in two further specific areas looked at measuring key areas of the EMRs and finally assessing the recording of outcomes compared to self-reported questionnaires. To our knowledge this is a unique approach to looking at both large and small health databases.

7.5 Overall Summary

Increasingly EMRs are being used to record patient data and this presents an enormous repository of clinical resource. By exploring this potential it can contribute to our understanding of patient needs, prescribing patterns, prevalence and incidence of disease and whether national policies for healthcare are delivering the benefits they set out to achieve. This study has finally concluded that EMRs represent different challenges depending on what needs to be extracted with structured data easier to extract than narrative (or free text) with the most difficult stored being imagery contained data such as pdf (portable document format) charts and scans. This study has shown that it is possible to take a collection of differently generated EMRs and transform them into knowledge and for comparisons to take place and meaningful conclusions derived.

8 Appendices

8.1 Ethical approval for QOF/COPD study

SRC Feedback

Researcher Name: Prof. Richard Hubbard
Organisation: University of Nottingham
SRC Reference Number: 13-018
Date: 5th July 2013
Study title: The Respiratory Health of the Nation – using The Health Improvement Network to quantify the burden of respiratory disease in the UK.
Committee opinion: Approved

The following feedback has been supplied by the SRC.

Notes from the Chair:

Excellent responses – Approved.

We are pleased to inform that you can proceed with the study as this is now approved. CSD Medical Research will let the relevant Ethics committee know this study has been approved by the SRC.

Once the study has been completed and published, it is important for you to inform CSD Medical Research in order for us to advise the SRC and your reference number to be closed.

References to all published studies are added to our website enabling other researchers to become aware of your work. Copies of publication(s), where available, will be appreciated.

I wish you and your team all the best with the study progression.



Mustafa Dungarwalla
Research Associate

SRC
Scientific Review Committee

8.2 Respiratory questionnaire 2000

Respiratory Questionnaire

University of Nottingham - Lung Health Study

Respiratory Medicine Unit, City Hospital, Nottingham

Study no.

Subject ID

Study team

Your date of birth .....

Male/Female .....

I AM GOING TO ASK YOU SOME QUESTIONS. AT FIRST, THESE WILL BE MOSTLY ABOUT YOUR BREATHING. WHEREVER POSSIBLE, I WANT YOU TO ANSWER YES OR NO BY CIRCLING THE APPROPRIATE ANSWER..

Yes ***No***

WHEEZE AND TIGHTNESS IN THE CHEST

1. Have you had wheezing or whistling in your chest

at any time in the last 12 months? ***Yes*** ***No***

2. Have you woken up with a feeling of tightness

in your chest at any time in the last 12 months? ***Yes*** ***No***

SHORTNESS OF BREATH

3. Have you had an attack of shortness of breath that came on

during the day when you were at rest at any time in the

last 12 months? ***Yes*** ***No***

4. Have you been woken by an attack of shortness of breath

at any time in the last 12 months? ***Yes*** ***No***

5. Have you had an attack of shortness of breath that
came on FOLLOWING strenuous activity at any time
in the last 12 months? ***Yes*** ***No***

COUGH

6. Have you ever been woken by an attack of coughing
at any time in the last 12 months? ***Yes*** ***No***

7. Have you had a head cold, sore throat, 'flu or chest infection
in the last week? ***Yes*** ***No***

- IF 'YES' TO QUESTION 7, GO TO QUESTION 9

8. Have you had a head cold, sore throat, 'flu or chest
infection in the last 4 weeks? ***Yes*** ***No***

COUGH & PHLEGM FROM THE CHEST

9. Do you USUALLY cough first thing in the morning
in the winter? ***Yes*** ***No***

10. Do you USUALLY cough during the day, or at night,
in the winter? ***Yes*** ***No***

- IF 'NO' TO QUESTIONS 9 & 10, GO TO QUESTION 12

11. Do you cough like this most days for as much as

three months each year? ***Yes*** ***No***

12. Do you USUALLY bring up any phlegm from your chest

first thing in the morning in the winter? ***Yes*** ***No***

13. Do you USUALLY bring up any phlegm from your chest

during the day, or at night, in the winter? ***Yes*** ***No***

- IF 'NO' TO QUESTIONS 12 AND 13, GO TO QUESTION 15

14. Do you bring up phlegm like this on most days for as

much as 3 months each year? ***Yes*** ***No***

BREATHING

15. Do you ever have trouble with breathing? ***Yes*** ***No***

- IF 'NO' TO QUESTION 15, GO TO QUESTION 17

16. Do you have this trouble: Tick box ✓

1. Continuously, so that your breathing is never quite right? ☐
2. Repeatedly, but it always gets completely better? ☐
3. Only rarely? ☐

17. Are you disabled from walking by a condition

OTHER THAN heart or lung disease? ***Yes*** ***No***

If 'YES', please state condition.....

- IF 'YES', GO TO QUESTION 21

18. Are you troubled by shortness of breath when hurrying

on level ground or walking up a slight hill? ***Yes*** ***No***

19. Do you get short of breath walking with other people

of your own age on level ground? ***Yes*** ***No***

- IF 'NO', GO TO QUESTION 21

20. Do you have to stop for breath when walking

at your own pace on level ground? ***Yes*** ***No***

ASTHMA

21. Have you ever had asthma? ***Yes*** ***No***

- IF 'NO', GO TO QUESTION 25

22. Was this confirmed by a doctor? ***Yes*** ***No***

23. How old were you when you had your first attack of asthma?

Age in years.....

24. Have you had an attack of asthma in the last 12 months? ***Yes*** ***No***

HAYFEVER & ECZEMA

25a.Have you EVER had hay fever or other nasal allergies? ***Yes*** ***No***

- IF 'NO' TO QUESTION 25a, GO TO QUESTION 26a

25b. Have these allergies ONLY occurred during the Summer

months? ***Yes*** ***No***

26a. IN THE LAST YEAR, have you had an itchy skin

condition - by itchy we mean scratching or rubbing the skin? ***Yes*** ***No***

- IF 'NO' TO QUESTION 26a, GO TO QUESTION 29

26b. Have you had this itchy skin condition

IN THE LAST WEEK? ***Yes*** ***No***

27. How old were you when this skin condition began?

Under 2 ☐ 2-5 years ☐ 5-10 years ☐ Over 10 years ☐

Uncertain, but 'as a child' ☐

28. Has this skin condition ever affected the skin creases in the past

-by skin crease, we mean fronts of elbows, behind the knees,

fronts of ankles, around the neck, or around the eyes? ***Yes*** ***No***

29. IN THE LAST YEAR, have you suffered from dry skin

in general? ***Yes*** ***No***

30. IN THE LAST YEAR, have you suffered from any of the

follwing skin complaints (please circle all that apply) :

a.) Eczema ***Yes*** ***No***

b.) Dermatitis ***Yes*** ***No***

c.) Psoriasis ***Yes*** ***No***

31. Have you EVER had eczema or any kind of skin rash? ***Yes*** ***No***

32. Have you EVER had an itchy dry skin rash in the joints

or creases of your body, such as behind you knees or in front

of your elbows? ***Yes*** ***No***

- IF 'NO' TO BOTH QUESTIONS 31 & 32, GO TO QUESTION 42

33. Is there such a rash NOW? ***Yes*** ***No***

- IF 'YES' TO QUESTION 33, GO TO QUESTION 35

34. Have you had such a rash IN THE LAST 12 MONTHS? ***Yes*** ***No***

35. Has the rash tended to come and go? ***Yes*** ***No***

36. Has the rash ever oozed or become sore? ***Yes*** ***No***

37. Is the rash made worse by soap, bubble bath, wool or nylon clothing? ***Yes*** ***No***

38. Has anybody in your immediate family had a rash like this? ***Yes*** ***No***

39. Have you ever had an itchy,dry rash on the hands or feet? ***Yes*** ***No***

- IF 'NO' TO QUESTION 39, GO TO QUESTION 41

40.Was this only the feet? ***Yes*** ***No***

41. Does your skin become itchy when you are hot? ***Yes*** ***No***

YOUR FAMILY

42. Were you born in the United Kingdom? ***Yes*** ***No***

- IF 'YES' TO QUESTION 42, GO TO QUESTION 44

43. How many years have you lived in the UK? ***Years***.....

44. How many brothers do/did you have? .....

- IF '0' OR DON'T KNOW, GO TO QUESTION 47

45. How many OLDER brothers? .....

46. How many of your brothers (older or younger)

ever had asthma? .....

47. How many sisters do/did you have? .....

- IF '0' OR DON'T KNOW, GO TO QUESTION 50

48. How many OLDER sisters? .....

49. How many of your sisters (older or younger)

ever had asthma? .....

50. Did you mother ever have asthma? ***Yes*** ***No*** ***Don't know***

51. Did you mother smoke regularly during your childhood? ***Yes*** ***No*** ***Don't know***

52. Did your father ever have asthma? ***Yes*** ***No*** ***Don't know***

53. Did you father smoke regularly during your childhood? ***Yes*** ***No*** ***Don't know***

YOUR HOME

54. How many years have you lived in your present home? .....

55. How many years have you lived in the same town? .....

56. Does your home have any of the following forms of heating? Tick all that apply ✓

- Central heating ☐
- Ducted air heating ☐
- Open fire (coal/wood) ☐
- Gas fire ☐
- Electric heater ☐
- Paraffin heater ☐
- Other ☐

57. What kind of fuel do you mostly use for cooking? Tick one box ✓

Solid fuel (coal, coke, wood, etc) ☐

Gas ☐

Electricity ☐

Paraffin ☐

Other ☐

YOUR PETS

58a. Do you keep any of the following pets? Tick box ✓

A cat ☐

A dog ☐

Any birds ☐

58b. Have you avoided keeping any pets because of allergies? ***Yes*** ***No***

SMOKING HABIT

59. Have you ever smoked for as long as a year? ***Yes*** ***No***

(Enter Yes if at least 1 cigarette/day or 1 cigar/week for 1 year)

- IF 'NO' TO QUESTION 59, GO TO QUESTION 68

60. How old were you when you started smoking? .....

61. Do you smoke now (as of ONE MONTH AGO)? ***Yes*** ***No***

- IF 'NO' TO QUESTION 61, GO TO QUESTION 64

62. How much do you now smoke on average?

(For own-rolled cigarettes, 28g (1oz) tobacco/wk = 4 cigarettes/day)

Number of cigarettes per day	.....
Number of cigarillos per day	.....
Number of cigars per week	.....
Amount of pipe tobacco per week	.....g

63. Have you smoked within the last hour? ***Yes*** ***No***

64. Have you stopped or cut down smoking? ***Yes*** ***No***

- IF 'NO' TO QUESTION 64, GO TO QUESTION 67

65. How old were you when you stopped or cut down smoking? .....

66. ON AVERAGE, of the entire time you smoked (before you stopped or cut down) how much did you smoke?

(For own-rolled cigarettes, 28g (1oz) tobacco/wk = 4 cigarettes/day)

Number of cigarettes per day	.....
Number of cigarillos per day	.....
Number of cigars per week	.....
Amount of pipe tobacco per week	.....g

67. Do you or did you inhale the smoke? ***Yes*** ***No***

68. Have you been REGULARLY (i.e., on most days or nights) exposed to other people's tobacco smoke in the last 12 months? ***Yes*** ***No***

- IF 'NO' TO QUESTION 68, GO TO QUESTION 70

69. How many hours per day are you exposed to OTHER

PEOPLE'S tobacco smoke? .....hours

Please turn to page 7...

MEDICINES & INHALERS

70. Have you used any INHALED medicines to help your

breathing at any time in the last 12 months? ***Yes*** ***No***

- IF 'NO' TO QUESTION 70, GO TO QUESTION 71

BRONCHODILATOR INHALERS

Please look at this list of inhalers and indicate which you have used in the last

12 months:

Tick box(es) ✓

1. Salbutamol (Ventolin, Ventodisks, Salbuvent, Ventide) ☐
2. Terbutaline (Brycanyl) ☐
3. Ipratropium (Atrovent, Duovent) ☐
4. Salmeterol (Serevent) ☐
5. Eformoterol (Oxis, Foradil) ☐
6. Other bronchodilator ☐

PROPHYLACTIC INHALERS

1. Beclamethasone (Becotide,Becodisks,Becloforte,Ventide) ☐
2. Budesonide (Pulmicort) ☐
3. Betamethasone (Bextasol) ☐
4. Cromoglycate (Intal) ☐
5. Nedocromil (Tilade) ☐
6. Other prophylactic ☐

71. Have you used any pills, capsules, tablets or medicines

(other than inhaled medicines)for breathing problems or

hay fever at any time in the last 12 months? ***Yes*** ***No***

- IF 'NO' GO TO END

ORAL BETA AGONISTS

Please look at this list of drugs and indicate which you have used in the last

12 months:

Tick box(es) ✓

1. Salbutamol (Volmax, Asmaven, Salbulin, Salbuvent) ☐
2. Terbutaline (Bricanyl) ☐
3. Other ☐

ORAL METHYLXANTHINES

Please look at this list of drugs and indicate which you have used in the last

12 months:

Tick box(es) ✓

1. Aminophylline (Phyllocontin) ☐
2. Theophylline (Nuelin, Theo-dur) ☐
3. Other ☐

LEUKOTRIENE ANTAGONISTS

Please look at this list of drugs and indicate which you have used in the last

12 months:

Tick box(es) ✓

1. Zafirleukast (Accolate) ☐
2. Monteleukast (Singulair) ☐
3. Other ☐

ORAL/SYSTEMIC STEROIDS

Please look at this list of drugs and indicate which you have used in the last

12 months:

Tick box(es) ✓

1. Prednisolone (Deltacortil, Prednesol) ☐
2. Triamcinolone (Ledercort, Kenalog) ☐
3. Other ☐

ORAL ANTIHISTAMINE

Please look at this list of drugs and indicate which you have used in the last

12 months:

Tick box(es) ✓

1. Astemizole (Hismanal) ☐
2. Terfenadine (Triludan) ☐
3. Loratadine (Clarytin) ☐
4. Cetirizine (Zirtek) ☐
5. Other ☐

THE QUESTIONNAIRE IS NOW COMPLETE

THANK YOU FOR YOUR HELP

8.3 Respiratory questionnaire 2007

Respiratory Questionnaire

The University of Nottingham

This questionnaire is going to ask questions about health and lifestyle factors. .
Please complete the questions as shown below. **Please fill in all the questions, unless you are told to skip them.**

If you have difficulties filling in the questionnaire, please do not hesitate to phone and ask for help from one of the team members.

Below are some examples on how to fill in the boxes

For Letters A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

For Numbers 0 1 2 3 4 5 6 7 8 9

For other responses, please place a x in box as shown ☒

Study ID Number

First Name

Last Name

Date of birth / /

Respiratory Questionnaire

WHEEZE AND TIGHTNESS IN THE CHEST

	Yes	No
1. Have you had wheezing or whistling in your chest at any time in the last 12 months?	☐	☐
2. Have you woken up with a feeling of tightness in your chest at any time in the last 12 months?	☐	☐

SHORTNESS OF BREATH

	Yes	No
3. Have you had an attack of shortness of breath that came on during the day when you were at rest at any time in the last 12 months?	☐	☐
4. Have you been woken by an attack of shortness of breath at any time in the last 12 months?	☐	☐
5. Have you had an attack of shortness of breath that came on FOLLOWING strenuous activity at any time in the last 12 months?	☐	☐

COUGH

	Yes	No
6. Have you ever been woken by an attack of coughing at any time in the last 12 months?	☐	☐

COUGH & PHLEGM FROM THE CHEST

	Yes	No
7. Do you USUALLY cough first thing in the morning in the winter?	☐	☐
8. Do you USUALLY cough during the day, or at night, in the winter?	☐	☐

IF 'NO' TO QUESTIONS 7 & 8, GO TO QUESTION 10

	Yes	No
9. Do you cough like this most days for as much as three months each year?	☐	☐
10. Do you USUALLY bring up any phlegm from your chest first thing in the morning in the winter?	☐	☐
11. Do you USUALLY bring up any phlegm from your chest during the day, or at night, in the winter?	☐	☐

IF 'NO' TO QUESTIONS 10 AND 11, GO TO QUESTION 13

	Yes	No
12. Do you bring up phlegm like this on most days for as much as 3 months each year?	☐	☐

BREATHING

	Yes	No
13. Do you ever have trouble with breathing?	☐	☐

IF 'NO' TO QUESTION 13, GO TO QUESTION 15

14. Do you have this trouble:
   a. Continuously, so that your breathing is never quite right? ☐
   b. Repeatedly, but it always gets completely better? ☐
   c. Only rarely? ☐

	Yes	No
15. Are you disabled from walking by a condition OTHER THAN heart or lung disease?	☐	☐

If 'YES', please state condition ☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐

IF 'YES', GO TO QUESTION 19

	Yes	No
16. Are you troubled by shortness of breath when hurrying on level ground or walking up a slight hill?	☐	☐
17. Do you get short of breath walking with other people of your own age on level ground?	☐	☐

IF 'NO', GO TO QUESTION 19

18. Do you have to stop for breath when walking at your own pace on level ground?	☐	☐

Respiratory Questionnaire

ASTHMA

	Yes	No
19. Have you ever had asthma?	☐	☐
IF 'NO', GO TO QUESTION 23		
20. Was this confirmed by a doctor?	☐	☐
21. How old were you when you had your first attack of asthma? ☐☐ years old		
	Yes	**No**
22. Have you had an attack of asthma in the last 12 months?	☐	☐

HAYFEVER & ECZEMA

	Yes	No
23. .Have you EVER had hayfever or other nasal allergies?	☐	☐
IF 'NO' TO QUESTION 23, GO TO QUESTION 25		
24. Have these allergies ONLY occurred during the Summer months?	☐	☐
25. Have you ever had eczema or a skin condition affecting the skin creases (by skin creases, we mean fronts of elbows, behind the knees, fronts of ankles, around the neck, or around the eyes)?	☐	☐
IF 'NO' TO QUESTION 25, GO TO QUESTION 29		
	Yes	**No**
26. Do you have current signs of eczema?	☐	☐
27. In the last year have you had eczema?	☐	☐
28. Did your eczema start as a child?	☐	☐

SMOKING HABIT

	Yes	No
29. Have you ever smoked for as long as a year?	☐	☐

(Enter Yes if at least 1 cigarette/day or 1 cigar/week for 1 year)

IF 'NO' TO QUESTION 29, GO TO QUESTION 36

30. How old were you when you started smoking? ☐☐

	Yes	No
31. Do you smoke now (within the last month)?	☐	☐

IF 'NO' TO QUESTION 31, GO TO QUESTION 34

32. How much do you now smoke on average?
(For own-rolled cigarettes, 28g (1oz) tobacco/wk = 4 cigarettes/day)

Number of cigarettes per day ☐☐☐

Number of cigars per week ☐☐☐

Amount of pipe tobacco per week ☐☐☐ in grams

	Yes	No
33. Have you stopped or cut down smoking?	☐	☐

IF 'NO' TO QUESTION 33, GO TO QUESTION 36

34. How old were you when you stopped or cut down smoking? ☐☐

35. ON AVERAGE, during the time you smoked (before you stopped or cut down) how much did you smoke?
(For own-rolled cigarettes, 28g (1oz) tobacco/wk = 4 cigarettes/day)

Number of cigarettes per day ☐☐☐

Number of cigars per week ☐☐☐

Amount of pipe tobacco per week ☐☐☐ in grams

	Yes	No
36. Have you been REGULARLY (i.e., on most days or nights) exposed to other people's tobacco smoke in the last 12 months?	☐	☐

IF 'NO' TO QUESTION 36, GO TO QUESTION 38

37. How many hours per day are you exposed to OTHER PEOPLE'S tobacco smoke? ☐☐

MEDICINES & INHALERS

	Yes	No
38. Have you used any INHALED medicines to help your breathing at any time in the last 12 months?	☐	☐

IF 'NO' TO QUESTION 38, GO TO QUESTION 40

39. Please look at this list of inhalers and indicate which you have used in the last 12 months:

RELIEVER INHALERS

1. Salbutamol (Ventolin, Ventodisks, Salbuvent) ☐
2. Terbutaline (Bricanyl) ☐
3. Ipratropium (Atrovent, Duovent) ☐
4. Combivent ☐
5. Other reliever please list ☐ _____

PREVENTER INHALERS

1. Beclometasone (Becotide,Becodisks,Becloforte,Ventide) ☐
2. Budesonide (Pulmicort) ☐
3. Fluticasone (Flixotide) ☐
4. Cromoglycate (Intal) ☐
5. Eformoterol (Oxis, Foradil) ☐
6. Mometasone (Asmanex, Twisthaler) ☐
7. Nedocromil (Tilade) ☐
8. Salmeterol (Serevent) ☐
9. Tiotropium (Spiriva) ☐
10. Other preventer inhalers, please list ☐ _____

COMBINATION INHALERS

1. Seretide (Serevent/flixotide) ☐
2. Symbicort (formoterol/budesonide) ☐
3. Other combination inhalers, please list ☐

_____

	Yes	No
40. Have you used any pills, capsules, tablets or medicines (other than inhaled medicines) for breathing problems or hay fever at any time in the last 12 months? ***IF 'NO' GO TO QUESTION 42***	☐	☐

41. Please look at this list of drugs and indicate which you have used in the last 12 months:

ORAL BETA AGONISTS

1. Salbutamol (Volmax, Asmaven, Salbulin, Salbuvent) ☐
2. Terbutaline (Bricanyl) ☐
3. Other ☐

ORAL METHYLXANTHINES

1. Aminophylline (Phyllocontin) ☐
2. Theophylline (Nuelin, Theo-dur) ☐
3. Other ☐

LEUKOTRIENE ANTAGONISTS

1. Zafirlukast (Accolate) ☐
2. Montelukast (Singulair) ☐
3. Other ☐

ORAL/SYSTEMIC STEROIDS

1. Prednisolone (Deltacortil, Prednesol) ☐
2. Other ☐

42. What do you generally use as a pain medication? **(Cross only one)**

Paracetamol/ Panadol	☐	Mefenamic acid	☐
Nurofen/Ibuprofen/Advil/Bufren	☐	Indometacin	☐
Aspirin	☐	Naproxen	☐
Diclofenac sodium	☐	Other, Please list	☐

_____

43. In the last year, how often on average have you taken Paracetamol/ Panadol?

Never	☐	One day a week	☐
Once a month or less	☐	2-6 days a week	☐
More than once a month but less than once a week	☐	Everyday	☐

44. How many Paracetamol/ Panadol do/did you take per day, on the days you take it?

1 per day	☐	3-8 per day	☐
2 per day	☐	>8 day	☐
Don't use	☐		

45. Does the above amount reflect your Paracetamol/ Panadol use over the past 15 years?

Yes ☐ No ☐

46. In the past 15 years on average, do you think you have used more or less Paracetamol/ Panadol than stated above?

More ☐ Less ☐ Similar amount ☐

47. In the past 15 years, how often on average have you taken Paracetamol/ Panadol for each of the symptoms named?

	Never	Daily	2-6 days/week	1 day/week	Less than weekly	Less than monthly
Headache	☐	☐	☐	☐	☐	☐
Fever	☐	☐	☐	☐	☐	☐
Rheumatic pain	☐	☐	☐	☐	☐	☐
Toothache	☐	☐	☐	☐	☐	☐
Cough	☐	☐	☐	☐	☐	☐
Shortness of breath	☐	☐	☐	☐	☐	☐
Running nose or sneezing	☐	☐	☐	☐	☐	☐
Lung congestion	☐	☐	☐	☐	☐	☐
Common cold	☐	☐	☐	☐	☐	☐
Other (specify_____)	☐	☐	☐	☐	☐	☐

48. Do you avoid taking Paracetamol/ Panadol? Yes ☐ No ☐

If yes, why?_____

49. In the last year, how often have you taken Ibuprofen/Advil/Bufren?

Never	☐	One day a week	☐
Once a month or less	☐	2-6 days a week	☐
More than once a month but less than once a week	☐	Everyday	☐

50. How many Nurofen/Ibuprofen/Advil/Bufren do/did you take per day, on the days you take it?

1 per day	☐	3-8 per day	☐
2 per day	☐	>8 day	☐
Don't use	☐		

51. Does the above amount reflect your Nurofen/Ibuprofen/Advil/Bufren use over the past 15 years?
Yes ☐ No ☐

52. In the past 15 years on average, do you think you have used more or less Nurofen/Ibuprofen/Advil/Bufren than stated above?

More ☐ Less ☐ Similar amount ☐

53. In the past 15 years, how often have you taken Nurofen/Ibuprofen/Advil/Bufren for each of the symptoms named?

	Never	Daily	2-6 days/week	1 day/week	Less than weekly	Less than monthly
Headache	☐	☐	☐	☐	☐	☐
Fever	☐	☐	☐	☐	☐	☐
Rheumatic pain	☐	☐	☐	☐	☐	☐
Toothache	☐	☐	☐	☐	☐	☐
Cough	☐	☐	☐	☐	☐	☐
Shortness of breath	☐	☐	☐	☐	☐	☐
Running nose or sneezing	☐	☐	☐	☐	☐	☐
Lung congestion	☐	☐	☐	☐	☐	☐
Common cold	☐	☐	☐	☐	☐	☐
Other (specify_____)	☐	☐	☐	☐	☐	☐

54. Do you avoid taking Ibuprofen/Advil/Bufren? Yes ☐ No ☐

If yes, why?_____

55. In the last year, how often have you taken Aspirin?

Never ☐ One day a week ☐
Once a month or less ☐ 2-6 days a week ☐
More than once a month but less than once a week ☐ Everyday ☐

56. How many Aspirin do/did you take per day, on the days you take it?

1 per day ☐ 3-8 per day ☐
2 per day ☐ >8 day ☐
Don't use ☐

57. Does the above amount reflect your Aspirin use over the past 15 years?
Yes ☐ No ☐

58. In the past 15 years on average, do you think you have used more or less Aspirin than stated above?

More ☐ Less ☐ Similar amount ☐

59. In the past 15 years, how often have you taken Aspirin for each of the symptoms named?

	Never	Daily	2-6 days/week	1 day/week	Less than weekly	Less than monthly
Headache	☐	☐	☐	☐	☐	☐
Fever	☐	☐	☐	☐	☐	☐
Rheumatic pain	☐	☐	☐	☐	☐	☐
Toothache	☐	☐	☐	☐	☐	☐
Cough	☐	☐	☐	☐	☐	☐
Shortness of breath	☐	☐	☐	☐	☐	☐
Running nose or sneezing	☐	☐	☐	☐	☐	☐
Lung congestion	☐	☐	☐	☐	☐	☐
Common cold	☐	☐	☐	☐	☐	☐
Other (specify_____)	☐	☐	☐	☐	☐	☐

60. Do you avoid taking Aspirin?

Yes ☐ No ☐

If yes, why?_____

THIS QUESTIONNAIRE IS NOW COMPLETE

THANK YOU FOR YOUR HELP

8.4 Ethical approval for the Gedling Cohort Study

NHS
National Research Ethics Service

North Nottinghamshire Research Ethics Committee
1 Standard Court
Park Row
Nottingham
NG1 6GN

Telephone: 0115 9123344 ext. 39368
Facsimile: 0115 9123300

23 December 2008

Dr Tricia McKeever
Wellcome Trust Fellow in Clinical Epidemiology
Division of Respiratory Medicine
Clinical Sciences Building
Nottingham University Hospitals NHS Trust
City Hospital
Hucknall Road
Nottingham, NG5 1PB

Dear Dr McKeever

Study title: **The effects of diet in the aetiology and natural history of asthma and chronic obstructive pulmonary disease – Gedling follow up 2005**
REC reference: **05/Q2402/18**

This study was given a favourable ethical opinion by the Committee on 18 August 2005.

It is a condition of approval by the Research Ethics Committee that the Chief Investigator should submit a progress report for the study 12 months after the date on which the favourable opinion was given, and then annually thereafter. To date, the Committee has not yet received the annual progress report for this year, which was due on 18 August 2008 despite a reminder letter being sent 1 October 2008. It would be appreciated if you could complete and submit the report by no later than 23 January 2009.

Guidance on progress reports and a copy of the standard NRES progress report form is available at
http://www.nres.npsa.nhs.uk/applicants/after-ethical-review/progress-reports

There is also guidance on declaring the end of the study at
http://www.nres.npsa.nhs.uk/applicants/after-ethical-review/endofproject

05/Q2402/18	Please quote this number on all correspondence

Yours sincerely,

P Wheat

Ms Trish Wheat
Committee Co-ordinator

E-mail: trish.wheat@nottspct.nhs.uk

Copy to: R&D office for NHS care organisation at lead site – NUH (via email)
Mr Paul Cartledge – University of Nottingham

This Research Ethics Committee is an advisory committee to East Midlands Strategic Health Authority
The National Research Ethics Service (NRES) represents the NRES Directorate within the National Patient Safety Agency and Research Ethics Committees in England

WPH 1370

8.5 Letter of access to primary care practices

Nottinghamshire Healthcare NHS
NHS Trust

Positive about integrated healthcare

Nottinghamshire Healthcare NHS Trust
Institute of Mental Health
University of Nottingham Innovation Park
Triumph Road
Nottingham
NG7 2TU

08/09/2014

Vishal Basra
PhD Researcher
UKCTCS
University of Nottingham
Clinical Sciences Building
City Hospital
Hucknall Road
Nottingham
NG5 1PB

Dear Vishal,

Project Title: The effects of diet in the aetiology and natural history of asthma and chronic obstructive pulmonary disease. Gedling follow-up 2005
REC ref: 05/Q2402/18

Letter of access for research issued by Nottinghamshire Healthcare NHS Trust on behalf of Nottinghamshire County CCGs and Nottingham City GP Practices

This letter confirms your right of access to conduct research through **Nottinghamshire County CCGs and Nottingham City GP Practices** for the purpose and on the terms and conditions set out below. This right of access commenced on **30/09/2014** and ends on **30/09/2015** unless terminated earlier in accordance with the clauses below.

You have a right of access to conduct such research as confirmed in writing in the letter of permission for research from this NHS organisation. Please note that you cannot start the research until the Principal Investigator for the research project has received a letter from us giving permission to conduct the project.

The information supplied about your role in research **Nottinghamshire County CCGs and Nottingham City GP Practices** has been reviewed and you do not require an honorary research contract with this NHS organisation. We are satisfied that such pre-engagement checks as we consider necessary have been carried out.

You are considered to be a legal visitor **Nottinghamshire County CCGs and Nottingham City GP Practices** premises. You are not entitled to any form of payment or access to other benefits provided by this NHS organisation to employees and this letter does not give rise to any other relationship between you and this NHS organisation, in particular that of an employee.

While undertaking research through **Nottinghamshire County CCGs and Nottingham City GP Practices** you will remain accountable to your employer **University of**

Nottingham but you are required to follow the reasonable instructions of the relevant manager in this NHS organisation or those given on her/his behalf in relation to the terms of this right of access.

Where any third party claim is made, whether or not legal proceedings are issued, arising out of or in connection with your right of access, you are required to co-operate fully with any investigation by this NHS organisation in connection with any such claim and to give all such assistance as may reasonably be required regarding the conduct of any legal proceedings.

You must act in accordance with **Nottinghamshire County CCGs and Nottingham City GP Practices** policies and procedures, which are available to you upon request, and the Research Governance Framework.

You are required to co-operate with **Nottinghamshire County CCGs and Nottingham City GP Practices** in discharging its duties under the Health and Safety at Work etc Act 1974 and other health and safety legislation and to take reasonable care for the health and safety of yourself and others while on **Nottinghamshire County CCGs and Nottingham City GP Practices** premises. You must observe the same standards of care and propriety in dealing with patients, staff, visitors, equipment and premises as is expected of any other contract holder and you must act appropriately, responsibly and professionally at all times.

You are required to ensure that all information regarding patients or staff remains secure and *strictly confidential* at all times. You must ensure that you understand and comply with the requirements of the NHS Confidentiality Code of Practice (http://www.dh.gov.uk/assetRoot/04/06/92/54/04069254.pdf) and the Data Protection Act 1998. Furthermore you should be aware that under the Act, unauthorised disclosure of information is an offence and such disclosures may lead to prosecution.

You should ensure that, where you are issued with an identity or security card, a bleep number, email or library account, keys or protective clothing, these are returned upon termination of this arrangement. Please also ensure that while on the premises you wear your ID badge at all times, or are able to prove your identity if challenged. Please note that this NHS organisation accepts no responsibility for damage to or loss of personal property.

We may terminate your right to attend at any time either by giving seven days' written notice to you or immediately without any notice if you are in breach of any of the terms or conditions described in this letter or if you commit any act that we reasonably consider to amount to serious misconduct or to be disruptive and/or prejudicial to the interests and/or business of this NHS organisation or if you are convicted of any criminal offence. As from 26 July 2010, your HEI employer may initiate your Independent Safeguarding Authority (IISA) registration (where applicable), and thereafter, will continue to monitor your ISA registration status via the on-line ISA service. Should you cease to be ISA-registered, this letter of access is immediately terminated. Your employer will immediately withdraw you from undertaking this or any other regulated activity. You MUST stop undertaking any regulated activity.

Your substantive employer is responsible for your conduct during this research project and may in the circumstances described above instigate disciplinary action against you.

Nottinghamshire County CCGs and Nottingham City GP Practices will not indemnify you against any liability incurred as a result of any breach of confidentiality or breach of

the Data Protection Act 1998. Any breach of the Data Protection Act 1998 may result in legal action against you and/or your substantive employer.

If your current role or involvement in research changes, or any of the information provided in your Research Passport changes, you must inform your employer through their normal procedures. You must also inform your nominated manager in this NHS organisation.

Yours sincerely



Shirley Mitchell
Head of Research Management and Governance

cc: University of Nottingham HR department

8.6 Sample letter to Practices

The University of Nottingham

UNITED KINGDOM • CHINA • MALAYSIA

School of Medicine
Division of Epidemiology and Public Health
Clinical Sciences Building, Phase II
City Hospital
Hucknall Road
Nottingham
NG5 1PB
t:+44 (0)115 82 31340
f:+44 (0)115 82 31337
www.nottingham.ac.uk/medicine
Head of Division: Professor Richard Hubbard

[Practice Name]

[Date:]

[sample]

Dear Doctor,

In 1991 the above patients who are currently registered with your practice participated in a study of the relation between diet and lung disease, particularly asthma and COPD, carried out in a random sample of adult residents in the Gedling district of Nottingham. Over 2500 people took part in the original study, and we have since maintained contact with and been able to collect information on the incidence of disease on subsequent occasions from many of them.

In 2007, these patients were again contacted and participated in a questionnaire based study investigating individual dietary fatty acids may play an important role in determining the risk of airways disease. We are hoping, with patient permission, to extend our research to investigate the relation between diet and other common medical diseases including cardiovascular disease, cancer, diabetes, rheumatoid arthritis and liver disease. The above patients have granted us access to a copy of their GP record, which will allow us to expand our research within this cohort.

We are hoping to arrange a time, at your convenience, when we can come extract the electronic medical record and obtain a copy of the paper record. Please free to call me to arrange this, otherwise I shall be in touch in the near future to arrange a suitable date for me to visit the practice and collect the information.

This project is funded by the ESRC and has been approved by the North Nottinghamshire Local Research Ethics Committee. If you have any questions about the study please contact Vishal Rai Basra at Nottingham City Hospital (details below).

Thank you for your help.

Yours sincerely,

Vishal R. Basra Dr Tricia McKeever Professor John Britton

8.7 Extraction software on compact disc attached

Please find it on CD attached.

8.8 Form and Dosage Lists

Forms List:

2\-piece\sostomy\ssystem	bar
2\-piece\sostomy\ssystem\sadhesive\srings	barrier
2\-piece\sostomy\ssystem\sbags	barrier\scream
2\-piece\sostomy\ssystem\sbarrier	barrier\sfilm\spump\sspray
2\-piece\sostomy\ssystem\sbase\splates	barrier\swipes
2\-piece\sostomy\ssystem\sbelt\sadaptors	base\splate
2\-piece\sostomy\ssystem\scaps	bath\sadditive
2\-piece\sostomy\ssystem\scontrol\sdevice	bath\semollient
2\-piece\sostomy\ssystem\sflanges	bath\semulsion
2\-piece\sostomy\ssystem\splugs	bath\soil
2\-piece\sostomy\ssystem\spouches	bath\ssolution
2\-piece\sostomy\ssystem\sstoma\scaps	bavettes
2\-piece\sostomy\ssystem\swafers	bed\spad
absorbent\spad	below\sknee\sstocking
absorbent\sperforated\sdressing\swith\sadh	belt
esive\sborder	biscuits
adaptor	bladder\sinstillation
adaptor\sand\stube	bladder\sirrigation
adaptor\spack	body\ssuit
additive	bone\scement
adhesive	boot
adhesive\saerosol	border\sdressing
adhesive\sdisc	bord\ssof\ssilico\ss\-a
adhesive\sdressing	bottle
adhesive\sfilm\sdressing\swith\sabsorbent\	bottleliquid
spad	bread
adhesive\splaster	breadmix
adhesive\sremover	breadsticks
adhesive\ssolution	bread\splus\ssoya\sbran
adhesive\sstrip	bread\srolls
adhesive\stape	breath\-actuated\saerosol\sinhaler
aerosol	breath\-actuated\sdry\spowder\sinhaler
aerosol\sinhaler	bridge
aerosol\sspray	briefs
aerosol\ssublingual\sspray	brown\sbread
alcoholic\slotion	brown\sflour
alcoholic\ssoln	brown\srice\sbread
alcoholic\ssolution	brush
ampoules	buccal\stablets
anal\sirrigation\ssystem	bun
anellini	bung
anklet	burger\sbuns
appliance	button\swashout\sset
application	caecostomy\swashout\sset
application\scream	cake
application\senema	cakemix
application\sfoam\senema	cannelloni
applicator	cap
aqueous\sgel	caplets
aqueous\snasal\sspray	capsule
arm\ssleeve	capsules
atomiser	capsules\splus\scream
baby\sand\sinfant\ssuspension	capsules\splus\sHandiHaler
bag	capsules\splus\sinhaler
bags	capsules\s\(gastro\-resistant\)
baguettes	cartridge
bag\scover	cartridges
bag\scovers	cassette
bag\splus\stap\soutlet	catheter
bag\ssupport	catheter\smaintenance\ssolution
bag\s\+\stop	catheter\splus\sbag
bakemix	catheter\splus\sbag\splus\swater
balm	crm
bandage	crushable\stablets
catheter\svalve	crystals
caviject	cup
cereal	cushion

cervical\scollar	cutting\sneedle
cervical\sgel	cuvettes
CFC\-free\sinhaler	cylinder
CFC\-free\sspray	dental\sgel
chesty\scough\slinctus	dental\slacquer
chewable\scapsules	dental\srinse
chewable\stablets	deodorant
chewing\sgum	deodorant\sgel
chocolate\schip\scookies	deodorant\spowder
clamp	deodorant\ssolution
cleanser	deodorant\sspray
clickhaler	dessert
clip	device
collecting\sbag	diabetic\slinctus
collodion\sBP	diaphragm
colobelt	digestive\sbiscuits
colostomy\sbag	dilatation\scatheter
colostomy\sbelt	dilator
colostomy\splug	director
colostomy\sset	disc
colotip	diskhaler
colourless\scream	disks
comb	disks\splus\sdisk\sinhaler
compound\slinctus	disks\srefill
compound\smouthwash\stablets	dispenser
compression\sbandage	dispersible\sgastro\-resistant\stablets
compression\shosiery	dispersible\stablet
cone	dispersible\stablets
cone\stop	ditalini
connecting\stubing	double\ssided\splaster
connector	double\sspring\struss
container	douche\splus\sfittings
convertor	drain
cookies	drainage\sbag
corn\smix	dressing
corn\srings	dressings
cot	dribbling\sbag
cotton	drink
country\sloaf	dropper
cous\scous	drops
cover	dry\spowder\sinhaler
covering\scream	dry\spowder\sspray
cover\scream	Duopak
crackerbread	durable\sbarrier\scream
crackers	Durules
cracker\stoast	dusting\spowder
cre	ear\/eye\/nose\sdrops
cream	ear\sdrops
cream\sbefore\spessary	ear\sointment
cream\splus\spowder	ear\sspray
cream\srouge	effervescent\sgranules
cream\sshampoo	effervescent\sgranules\ssachets
cream\s\+\sdressing	effervescent\spowder\ssachets
crispbread	effervescent\stablets
crisp\srolls	egg\sreplacement
	egg\swhite\sreplacer
	elastic\sband\struss
	electrodes
	frame
elixir	funnel
emollient	fusilli\spasta
emollient\scream	gargle
Emulgel	garment
emulsion	gastro\-resistant\scapsules
enema	gastro\-resistant\sgranules
enteral\ssolution	gastro\-resistant\smodified\-release\scapsules
enteric\scoated\stablets	gastro\-resistant\stablets
evohaler	gauze
expectorant	gel
eye/ear/nose\sdrops	

eye/ear\sdrops	gelthix\scaps
eye/ear\sointment	gel\-forming\seye\sdrops
eye\sbath	gel\-forming\ssolution
eye\sdro	gel\sdressing
eye\sdrops	gel\skit
eye\sgel	gel\splus\sdressing
eye\sirrigation\ssolution	gel\splus\socclusive\sdressing
eye\sointment	general\spurpose\stype\sa\slancet
eye\sshade	gf\swf\spenne
eye\ssolution	gf\swf\sspaghetti
eye\swash	girdle
facepiece	gloves
face\smask	gluten\-free\sbread\smix
faecal\scollector\sbags	gluten\-free\scrackers
fatty\sointment	gluten\-free\scrisp\sbreads
felt	gluten\-free\spasta\scorn\slasagne
female\scatheter	gluten\-free\spasta\scorn\sspirals
female\scatheter\splus\sbag	gluten\-free\sself\sraising\sflour
female\scatheter\splus\sbag\splus\swater	gluten\-free\swhite\sbread\ssliced
fettucini	gluten\sfree
fibre\sloaf	gluten\sfree\sbiscuits
fibre\smix	gluten\sfree\sbread\srolls
film	gluten\sfree\scake\smix
film\-coated\stablets	gluten\sfree\scrackers
film\sdressing	gluten\sfree\scrispbread
filter	gluten\sfree\sdigestive\sbiscuits
filter\scassette	gluten\sfree\sfibre\sbread\srolls
finger\sstall	gluten\sfree\sfibre\sloaf\ssliced
finishing\spowder	gluten\sfree\sfibre\srolls
fixing	gluten\sfree\sflour\smix\sfor\scooking
fixing\spowder	gluten\sfree\sfresh\sbrown\sloaf\ssliced
flange	gluten\sfree\sfresh\swhite\sloaf\ssliced
flange\sextender	gluten\sfree\smatzo\scrackers
flange\splus\ssheath	gluten\sfree\smix
flatus\spatch	gluten\sfree\spart\sbaked\sfibre\srolls
flour	gluten\sfree\spart\sbaked\swhite\srolls
flour\smix	gluten\sfree\spasta\sfusilli
foam	gluten\sfree\spasta\slasagne
foam\saerosol\senema	gluten\sfree\spasta\spenne
foam\saerosol\sspr	gluten\sfree\spasta\sspaghetti
foam\sdressing	gluten\sfree\spastry\smix
foam\spad	gluten\sfree\spizza\sbase
foam\sring	gluten\sfree\sshortbread\sbiscuits
foot\sbag	gluten\sfree\sshortcake\srings
foundation	gluten\sfree\ssweet\sbiscuits
	gluten\sfree\swheat\sfree\sbread\smix
	gluten\sfree\swheat\sfree\sfibre\srolls
	gluten\sfree\swheat\sfree\slunch\srolls
	gluten\sfree\swheat\sfree\spastry\smix
	gluten\sfree\swheat\sfree\swhite\srolls
	gluten\sfree\swhite\sbread
	injection\sgel
gluten\sfree\swhite\sloaf\ssliced	injection\skit
gluten\sfree\swhite\slong\srolls	injection\spen
gluten\sfree\swhite\smix	injection\splus\sdiluent
gluten\sfree\swhite\srolls	injection\splus\stransfer\sneedle
granules	injection\spowder\,\sdiluent\sand\ssyringe
granules\sfor\ssuspension	injection\spowder\sfor\sreconstitution\
granules\splus\sdiluent	injection\spowder\splus\sdiluent\
granules\splus\ssolution	injection\spowder\s\+\sdiluent\s\+\skit\
granules\splus\svehicle	injection\srefills
grey\stoner	injection\ssolution
haemofiltration\ssolution	injection\ssuspension
half\scircle	injection\svial
hand\srub	inj\ssoln
high\sfibre\scrackers	inserts
holder	insufflator
holster	insufflator\scartridges
holster\sbag	insulin\spen
hydrogel\sdressing	intradermal\sinjection
hypodermic\sinsulin\sneedles\sfor\spre\-	intramuscular\sinjection\spdr\sfor\srecon\

filled\s\/\sreusable\spen\sinjectors	intranasal\ssolution
ileostomy\sappliance	intrathecal\sinjection
ileostomy\sbag	intrauterine\ssystem
ileostomy\sset	intravenous\sinfusion
imitation\srice	intravenous\sinfusion\sconcentrate
implant	intravenous\sinfusion\spdr\sfor\srecon\
implant\spre\-filled\ssyringes	intravenous\sinfusion\splus\sbuffer
implant\ssafesystem\spre\-filled\ssyringes	intravenous\sinfusion\spowder\s\+\sdiluent\
impregnated\sgauze	intravenous\sinjection
incontinence\sbelt	intravenous\ssolution
incontinence\spouch	intra\-articular\sinjection
incontinence\sset	intra\-muscular\sinjection
incontinence\ssheath	intra\-ocular\sinjection
infant\ssuppositories	intra\-ocular\sinjection\sprefilled\ssyringe
infusion	intra\-ocular\sirrigation
infusion\sconcentrate	intra\-ocular\ssolution
infusion\sconcentrate\sdr\splus\sdiluent	intra\samniotic\sinjection
infusion\sconcentrate\spdr\sfor\srecon\	intra\sarticular/intramuscular\sinjection
infusion\skit	intra\sarticular/\sintradermal\sinjection
infusion\spack	irrigation
infusion\splus\sdiluent	irrigation\ssolution
infusion\spowder\sfor\sreconstitution\	irrigation\ssolution\saerosol\sspray
infusion\sset	IV\sdressings
infusion\ssolution	jelly
inhalation	jelly\smix
inhalation\scapsules	juice
inhalation\scapsules\splus\sinhaler	junior\scapsules
inhalation\scartridges\swith\sdevice	junior\slozenges
inhalation\spowder	kipper\sbag
inhalation\spowder\scapsules	kit
inhalator	kneecap
inhaler	lancets
inhaler\skit	laryngectomy\sprotector
inhaler\splus\sspacer	laryngectomy\sT\scollar
inhaler\srefill	lasagne
inhaler\ssolution	latex\stube
initial\streatment\sset	leggings
inj	leg\sbags
injection	leg\sbag\sand\ssleeve
injection\sampoules	leg\ssleeve
injection\scartridge	leg\sstrap
injection\scartridges	leg\sstraps
injection\scartridge\spowder\s\+\sdiluent\	leg\sulcer\swrap
injection\sconcentrate	mouth\sspray
	muco\-adhesive\sbuccal\stablets
limb\sprotector\sshort\sleg	multigrain\swhite\sbread\(sliced\)
linctus	multi\-dose\svial
liner	nail\slacquer
liners	nasal\saspirator
linguine\spasta	nasal\sdilator
liniment	nasal\sdrops
lint	nasal\sgel
lipocream	nasal\sointment
lipstick	nasal\sspray
lip\sprotector	nebuliser
liquid	nebuliser\sliquid
liquid\sand\spack	nebuliser\ssolution
liquid\sconcentrate	nebuliser\sunits
liquid\seye\sgel	needle
liquid\sgel	needles\sfor\spre-filled\s\/\sreusable\spen\sinjectors
loaf	needle\sclipping
lotion	nephrostomy\sbag
lotio\-gel	nephrostomy\sbelt
lozenge	nephrostomy\sdrainage\ssystem
lozenges	net\ssuspensory\ssystem
lubricant	night\sbag
lubricating\sjelly	night\sdrainage\sbag
macaroni	nose\sdrops
macaroni\spenne	nose\sgel
maintenance\sset	no\ssting\sbarrier\sfilm\sfoam\sapplicators
male\scatheter	

male\scatheter\splus\sbag	nylon\smonofilament\ssuture
male\scatheter\splus\sbag\splus\swater	nylon\sstud
mask	ocular\sinsert
masking\scream	oily\scream
material	oily\sinjection
matrigel\scapsules	ointment
medicated\splasters	ointment\s\&\ssuppositories
melt	ophthalmic\sapplicator\sstrips
melt\stablets	ophthalmic\sinsert
membrane	ophthalmic\ssolution
micro\-enema	oral\sdrops
milk	oral\semulsion
milkshake	oral\sgel
minicap	oral\sliquid
mini\-tabs	oral\slyophilisates
mix	oral\spaint
mixture	oral\spaste
modified\-release\sbuccal\stablets	oral\spowder
modified\-release\scapsules	oral\spowder\ssachets
modified\-release\seye\sdrops	oral\srinse
modified\-release\stablets	oral\ssolution
modified\spowder\sgluten\sfree	oral\sspray
modified\srelease\sbuccal\stablets	oral\ssuspension
modified\srelease\scapsules	oral\ssyringe
modified\srelease\seye\sdrops	orodispersible\stablets
modified\srelease\sgranules	oro\-dispersible\stablets
modified\srelease\sinjection	orthopaedic\spadding
modified\srelease\stablet	ostomy\sbelt
modified\srelease\stablets	oxygen\smask
moisturiser	oxygen\sset
monofilament	O\-rings
mousse	pack
mouthpiece	packing
mouthwash	pad
mouthwash\sand\sgargle	pad\sdressing
mouth\sadaptor	plaster
paediatric\scapsules	plate
paediatric\scatheter	plug
paediatric\scatheter\splus\sbag	pouch
paediatric\scatheter\splus\sbag\splus\swate	poultice
r	powder
paediatric\sdrops	powder\sand\ssolvent\sfor\ssolution\sfor\sinj
paediatric\selixir	ection\svials
paediatric\sinjection	powder\sand\ssolvent\sfor\ssuspension\sfor\s
paediatric\slinctus	injection\svials
paediatric\sliquid	powder\sfor\sprolonged\srelease\sinjection\ss
paediatric\smixture	uspension
paediatric\ssolution	powder\sfor\sreconstitution
paediatric\ssugar\sfree\ssuspension	powder\sfor\sreconstitution\splus\sdiluent
paediatric\ssuppositories	powder\sfor\ssolution\sfor\sinjection
paediatric\ssuspension	powder\sfor\ssolution\sfor\sinjection\svials
paediatric\ssyrup	powder\sfor\ssuspension
paint	powder\splus\senema\skit
panty\sbrief	powder\splus\skit
panty\sgirdle	powder\sspray
paper	prefilled\spen
part\sbaked\sfibre\sloaf	prefilled\ssyringe
part\sbaked\sfibre\srolls	prefilled\ssyringe\spowder\splus\sdiluent\
part\sbaked\slong\swhite\srolls	preloaded\sinjection\spen
part\sbaked\swhite\sloaf	protector
part\sbaked\swhite\srolls	pt\sbkd\slong\sroll\swht
pasta	pump\sspray
pasta\salphabets	pump\sspray\spump\sspray
pasta\srings	pump\ssublingual\sspray
pasta\sshells	pyjamas
pasta\sspirals	p\-fill
pasta\stwists	rat\stail\struss
paste	reagent
paste\splus\sdressings	reagent\sstrips
pastilles	reagent\stablets
pastry\smix	receiver

patch	rectal\sfoam
patches	rectal\sointment
patches\sand\stablets	rectal\ssolution
peak\sflow\smeter	Rectube
pellets	refill
penile\ssheath	refill\scanister
penis\sholder	refill\sinjection
penne\spasta	release\sunit
periodontal\sgel	reliever
peritoneal\sdialysis\ssolution	respirator\ssolution
pessaries	retainer
pessaries\splus\scream	retention\senema
pessary	ribbon\sdressing
pessary\splus\scream	rice
petites\sbaguettes	rice\sbread
pill	rice\scookies
pills	rice\slasagne
pipe	rice\sspaghetti
pizza\sbase	rigati
	ring
	rings
	rolls
	roll\-on
	rusks
	sachet
	sachets
	salt\sfree\sbread
	sausage\sbelt
	savory\sbites
savoy\sbiscuits	savoury\sbiscuits
scalp\sand\sskin\scleanser\ssolution	spaghetti\s\short\scut\
scalp\sapplication	Spansules
scalp\slotion	spare
scalp\sointment	spare\schamber
scalp\ssolution	speech\svalve
Scott\scatheter	spincaps
seal	spirals
sealant\ssolution	spirit
seal\-tight	spotstick
self\-aspirating\scartridge	spray
self\-raising\sspec\sdiet\sflr\/loaves	spray\sapplication
semi\-solid	spray\srefill
setting\spowder	spray\ssolution
shampoo	spread
sheath	spring\struss
sheath\sbag	sp\sdiet\sflr\/loaf\/
sheath\scollar	square
sheath\sholder	squash
sheet	stall
sheets	starter\spack
sheet\sdressing	sterile\sdressing\spack
shield	sterile\spowder
shirt	sterile\ssolution
shortbread	sterile\ssuspension
shortbread\sbiscuits	sterile\sswabs
shortcake\srings	stick
shorts	stockinette
showering\said	stocking
shower\semollient	stocking\sapplicator
silicone\sglue	stoma\sbridge
silk\ssuture	stoma\scap
single\sdose\sinjection\svial	stoma\scollars
single\sdose\sunit\seye\sdrops	stopcock
single\sdose\sunit\seye\sgel	strap
single\sdose\sunit\snasal\sdrops	straps
single\srat\stail\struss	strip
single\sspring\struss	strips
skin\sbarrier\sspray	stud
skin\scleanser\ssolution	subcutaneous\sinjection
skin\sprick\stesting\skit	sublingual\stablets
sleeve	sub\-

sliced\sbread	unit\svaccine\ssuspension\sfor\sinjection
sliced\scountry\sloaf	sugar\-free\ssuspension
sliced\sloaf	sugar\sfree\schewable\stablets
sliced\swhite\sloaf	sugar\sfree\sdispersible\stablets
sling	sugar\sfree\selixir
soap	sugar\sfree\sgranules
sock	sugar\sfree\slinctus
soluble\stablets	sugar\sfree\sliquid
solution	sugar\sfree\slozenges
solution\sfor\sinjection	sugar\sfree\smixture
solution\sfor\sinjection\spre\-filled\ssyringes	sugar\sfree\soral\ssolution
solution\sfor\sinjection\svials	sugar\sfree\spaediatric\selixir
soya\sbran\sbread	sugar\sfree\spaediatric\slinctus
spacer	sugar\sfree\spaediatric\ssyrup
spacer\sdevice	sugar\sfree\spowder
spacer\sdevice\s\&\smask	sugar\sfree\ssuspension
spacer\sinhaler	sugar\sfree\ssyrup
spaghetti	supplement
spaghetti\srings	support
	supports
	support\sbelt
	support\sgarment
suprapubic\scatheter	suppositories
surgical\sscrub	tinted\scream
surgical\ssolution	tissue
suspension	tissue\sadhesive
suspension\sfor\sinjection	tonic
suspension\sfor\sinjection\svials	toothpaste
suspension\ssachets	topical\sgel
suspensory\ssystem	topical\sliquid
Sustets	topical\sliquid\ssachets
sut	topical\spowder
suture	topical\ssolution
sutures	torso\svest
suture\+11mm\sneedle	transdermal\spatch
suture\+16mm\sneedle	transdermal\spatches
suture\+19mm\sneedle	transurethral\sdelivery\ssystem
suture\+24mm\sneedle	treatment\sset
suture\+26mm\sneedle	triple\spack
suture\+35mm\sneedle	trousers
suture\+60mm\sneedle	tube
suture\swith\sneedle	tubing
suture\\+12mm\sneedle	tubular\sbandage
suture\\+18mm\sneedle	unit
suture\\+37mm\sneedle	unit\sdose\sblisters
swab	unit\sdose\svial
swabs	urethral\sinserts
sweet\sbiscuits	urethral\ssticks
syr	urinal
syringe	urinal\sbag
syringe\splus\sneedle	urostbag
syrup	urostomy\sbag
system	urostomy\sset
system\sbaseplate	urostset
system\stesting\sstrips	vaccine
tab	vaccine\semulsion\sfor\sinjection
tablet	vaccine\sinjection
tablets	vaccine\soral\ssuspension
tablets\splus\sgranules	vaccine\ssolution\sfor\sinjection
tablets\sstarter\spack	vaccine\ssolution\sfor\sinjection\spre\-filled\ssyringes
tablets\s\&\spessaries	vaccine\ssuspension\sfor\sinjection
tablet\spack	vaccine\s\(live\,\soral\)\sdrops
tabs	vacuum\spump
tab\.	vaginal\scapsules
tab\stablets	vaginal\scleansing\skit
tagliatelle	vaginal\scream
tampons	vaginal\sdilator
tape	vaginal\sgel
tapioca\sbread	vaginal\sjelly
tea\sbiscuits	vaginal\sring
test\skit	

test\sstrips	vaginal\stablets
test\stube	velcro\sleg\sstraps
Tetrapak	vermicelli
therapy\scapsules	vest
thermometer	vial
thigh\sstocking	vials
throat\sspray	vial\splus\sprefilled\ssyringe
ties	vial\s\+dil\sin\ssyringe\sinjection
tights	vitrellae
tincture	volatile\sliquid
	vortex\smetered\sdose\sinhaler
	wadding
	wafers
	waistband
washout\sappliance	wash
wax	washer
wax\sstick	
webbing\sbelt	
white\sbread	
white\sbread\sbuns	
white\sflour	
white\smix	
white\srolls	
wipes	
wool	
wound\sprotector|wound\sdrainage\sbags	

Dosages

\d+.\d+\w+|\d).*	\d\dmicrogram
\d\.\dmg\s\+\s\dmg\s\+\s\d\dmg	(([0-9.\.\-]+)mg\/([0-9.\.\-]+)ml)
(([0-9.]+)mg\/([0-9.]+)mg)	\d\s\+\s\d%
([\d]+mg	\d\.\d+\s\%\,\s\d\d\sml
[\d]+mcg	((\d+)micrograms/(\d+)hours)
[\d]+ml	([\d]+\.[\d]+m[lgcg])
\d+micrograms/dose	([\d]+\.[\d]+%)
\d\d\.\dmicrogram	([\d]+\smg
\d\.\d\dg\/\dml	[\d]+\smg)
\d+\.\d+mg|\d+\.\d+ml|\d+\.\d+mcg	([\d]+\.[\d]+m[gcgl])
\d+\.\d+g|\d+\.\d+gm	([\d]+micrograms\s\+\s[\d]+micrograms)
\dmm/\d\d	([\d\.\%]+\s\+\s[\d\.\%]+)
\d\.\dmm/\d\d	([\d]+mg\/[\d]+ml)
\d\dmm/\d\d	([\d\.]+g\/[\d]+ml)
\d%	([\d]+mg)
\d\d%	(\d%\s\+\s\d%)
\d\.\d\d%	(\d\dmcg\/\d\d\shours)
\d\.\d%	(\d\.\dmg\/\d\dhours)
\d\-\d\d	(\dmg\/\d\dhours)
\d\d\dmicrogram	(\d\d\dmg\/ml)
\d\dmicrogram	(\dmg\/\d\dhours)
\dmicrogram	(\d\dmicrograms\s\+\s\d\d\dmicrograms)
(([0-9.\.\-]+)g\/([0-9.\.\-]+)ml)	(\dmg\/\d\dhours)
(([0-9.]+)mg\/([0-9.]+)ml)	(\dmg\/\dml)
(\d%\/([0-9.]+)%)	(\dmicrograms/hour)
([0-9.]+)cm\sx\s([0-9.]+)cm	(\d\d\d\s[iu]/ml)
([0-9.\.]+)cm\sx\s([0-9.\.]+)cm	([\d]+\s\[iu]ml)
(([0-9.]+)m[g|l|cg]\s\+\s([0-9.]+)m[g|l|cg])	([\d]+\sm[gcgl])
(([0-9.]+)m[g|l|cg]\s\+\s([0-9.]+)m[g|l|cg])	
(([0-9.]+)microgram[s]\s\+\s([0-9.]+)microgram[s])	
\d\d\d\sgrams	

8.9 Downloading EMR steps

Methods Used for Data Extraction: Steps for each type of system encountered so far:-

Emis LV

F5 Patient Search

SP: Summary Printout

C: Defined Summary

Do you wish to use this terminal's default format: N

Q: Full Summary

Include any letters attached to this patient record: Y

Store as default summary printout for this terminal: N

Do you wish:

A.Lloyd George size printout

B.A4 Standard size printout

C.Right hand side of FP10

D.11' Fanfold continuous

E.11 Fanfold continuous

Select Option (ABCDE):

Print B

09 Export to Text file

Emis PCS

Click on patient

Type in patient name

Double click on patient found on list

Click on H-Full History

Click on N-Print/Export > scroll to 'Defined'

Ensure all boxes are ticket under 'Print Options'

Sort options left as default: Date & Descending

Click on 'Export'

Under export type dialog box select Text (CSV)

Click ok

Choose directory drive to store data

Asks: 'Do you wish to export attachments for this patient as files to disk?'

Click 'Yes'

Automatically saves in the same directory as patient record

TPP systmOne

On the SystemOne splash screen choose > 'Select patient'

Type in surname and press select

Then go to Patient >

Patient record > Print patient record

'Asks to print from'

Make sure second selection button checked and type in date 01 01 1900

For the options that say how much attachments are available – Choose save to disk

For options say any letters – Choose save to disk

Choose and book appointments – Choose skip

Should ask to save to letters and attachments to hard drive

Word file should be generated of complete record

i-SOFT from Computer Sciences Corporation: The download of iSOFT EMRs was done by the practice staff themselves. They were requested to provide complete EMRs of a patient as if they would be moving to another practice. They were able to do this but only in portable document format (PDF).

8.10 TRUD Dictionary: Both Read version 2 and Clinical Terms Version 3 (CTV3)

This was the 15.0.0 release pack and covers over 150,000 clinical codes. Description of its dictionary lists used. Files used for checking Read/CTV3 codes and Read and CTV3 terms were the following:

- Terms2 text file
- V2termv3 text file
- Descrip text file

8.11 Clinical Codes and Asthma Drugs

This appendix gives the clinical codes used for asthma and COPD (Read version 2 and Clinical Terms version 3 codes) and asthma drugs code list. Asthma codes from Mukherjee et al's study (164):

Code	Term
H33..	Asthma
H330.	Extrinsic (atopic) asthma
H3300	Extrinsic asthma without status asthmaticus
H3301	Extrinsic asthma with status asthmaticus
H330z	Extrinsic asthma NOS
H331.	Intrinsic asthma
H3310	Intrinsic asthma without status asthmaticus
H3311	Intrinsic asthma with status asthmaticus
H331z	Intrinsic asthma NOS
H332.	Mixed asthma
H333.	Acute exacerbation of asthma
H334.	Brittle asthma
H335.	Chronic asthma with fixed airflow obstruction
H33z.	Asthma unspecified
H33z0	Status asthmaticus NOS
H33z1	Asthma attack
H33z2	Late-onset asthma
H33zz	Asthma NOS
1780.	Aspirin induced asthma
1781.	Asthma trigger - pollen
1782.	Asthma trigger - tobacco smoke
1783.	Asthma trigger - warm air
1784.	Asthma trigger - emotion
1785.	Asthma trigger - damp
1786.	Asthma trigger - animals
1787.	Asthma trigger - seasonal
1788.	Asthma trigger - cold air
1789.	Asthma trigger - respiratory infection
178A.	Asthma trigger - airborne dust
178B.	Asthma trigger - exercise
1O2..	Asthma confirmed
66YK.	Asthma follow-up
663..	Asthma monitoring
663N.	Asthma disturbing sleep
663N0	Asthma causing night waking
663N1	Asthma disturbs sleep weekly
663N2	Asthma disturbs sleep frequently
663O.	Asthma not disturbing sleep
663O0	Asthma never disturbs sleep
663P.	Asthma limiting activities
663P0	Asthma limits activities 1 to 2 times per month
663P1	Asthma limits activities 1 to 2 times per week

Code	Term
663P2	Asthma limits activities most days
663Q.	Asthma not limiting activities
663U.	Asthma management plan given
663V.	Asthma severity
663V0	Occasional asthma
663V1	Mild asthma
663V2	Moderate asthma
663V3	Severe asthma
663W.	Asthma prophylactic medication used
663d.	Emergency asthma admission since last appointment
663e.	Asthma restricts exercise
663f.	Asthma never restricts exercise
663h.	Asthma - currently dormant
663j.	Asthma - currently active
663m.	Asthma accident and emergency attendance since last visit
663n.	Asthma treatment compliance satisfactory
663p.	Asthma treatment compliance unsatisfactory
663q.	Asthma daytime symptoms
663r.	Asthma causes night symptoms 1 to 2 times per month
663s.	Asthma never causes daytime symptoms
663t.	Asthma causes daytime symptoms 1 to 2 times per month
663u.	Asthma causes daytime symptoms 1 to 2 times per week
663v.	Asthma causes daytime symptoms most days
663w.	Asthma limits walking up hills or stairs
663x.	Asthma limits walking on the flat
663y.	Number of asthma exacerbations in past year
66YJ.	Asthma annual review
66YE.	Asthma monitoring due
66Y5.	Change in asthma management plan
66Y9.	Step up change in asthma management plan
66YA.	Step down change in asthma management plan
66Yq.	Asthma causes night time symptoms 1 to 2 times per week
66Yr.	Asthma causes symptoms most nights
66Yp.	Asthma review using Roy College of Physicians three questions
8CMA0	Patient has a written asthma personal action plan
8B3j.	Asthma medication review
8793.	Asthma control step 0
8794.	Asthma control step 1
8795.	Asthma control step 2
8796.	Asthma control step 3
8797.	Asthma control step 4
8798.	Asthma control step 5
9OJ1.	Attends asthma monitoring
9OJA.	Asthma monitoring check done
173d.	Work aggravated asthma
663e0	Asthma sometimes restricts exercise
663e1	Asthma severely restricts exercise
66YC.	Absent from work or school due to asthma

Code	Term
66YK.	Asthma follow-up
66YP.	Asthma night-time symptoms
66YQ.	Asthma monitoring by nurse
66YR.	Asthma monitoring by doctor
66Ys.	Asthma never causes night symptoms
66Yu.	Number days absent from school due to asthma in past 6 month
8791.	Further asthma - drug prevent.
8CR0.	Asthma clinical management plan

Read codes for asthma, version 3

Includes version 2 codes above and the following

Code	Term
XE0YX	Asthma NOS
Xa9zf	Acute asthma
XE0YW	Asthma attack
X1020	Pollen asthma
X1026	Baker's asthma
X102Y	Cardiac asthma
Ua1AX	Brittle asthma
XaIOV	Asthma finding
XaINi	Number of times bronchodilator used in one week
XaINj	Number of times bronchodilator used in 24 hours
XaIww	Asthma trigger
XaLIm	Asthma trigger - respiratory infection
XaLIn	Asthma trigger - seasonal
XaLIr	Asthma trigger - animals
XaLJS	Asthma trigger - cold air
XaLJT	Asthma trigger - damp
XaLJU	Asthma trigger - emotion
XaObi	Asthma trigger - airborne dust
XaObj	Asthma trigger - exercise
XaObk	Asthma trigger - pollen
XaObl	Asthma trigger - tobacco smoke
XaObm	Asthma trigger - warm air
XaYja	Asthma trigger - wind
XaYpF	Asthma trigger - perfume
X101x	Allergic asthma
XE2Nb	Asthma monitored
XaLPE	Nocturnal asthma
X101t	Childhood asthma
XaIer	Asthma follow-up
XaInC	Suspected asthma
XaIuG	Asthma confirmed
X1027	Colophony asthma
XM1Xb	Asthma monitoring
X101u	Late onset asthma
XM0s2	Asthma attack NOS
X1025	Industrial asthma

Code	Term
X102D	Status asthmaticus
XE0YV	Status asthmaticus NOS
XE0YT	Non-allergic asthma
X1025	Occupational asthma
X102D	Acute severe asthma
X101z	Allergic asthma NEC
X1023	Drug-induced asthma
XaQHq	Asthma control test
Xa0lZ	Asthmatic bronchitis
Xa8Hn	Asthma control steps
XaIeq	Asthma annual review
XaLJT	Asthma trigger - damp
XaYja	Asthma trigger - wind
XaIRN	Asthma monitoring due
X1020	Hay fever with asthma
X1028	Grain worker's asthma
XE0YQ	Allergic atopic asthma
XaJFG	Aspirin-induced asthma
Xa1hD	Exacerbation of asthma
XE0YV	Status asthmaticus NOS
XaBU2	Asthma monitoring call
XaKdk	Work aggravated asthma
X1029	Sulphite-induced asthma
XaIIZ	Asthma daytime symptoms
XaObk	Asthma trigger - pollen
XaLJU	Asthma trigger - emotion
XaYpF	Asthma trigger - perfume
XaBU3	Asthma monitoring status
XE2Na	Asthma monitoring admin.
XaBAQ	Recent asthma management
XaIfK	Asthma medication review
XaDvK	Asthma - currently active
XaJtu	Referral to asthma clinic
X1021	Allergic non-atopic asthma
XaIoE	Asthma night-time symptoms
XaIu5	Asthma monitoring by nurse
XaDvL	Asthma - currently dormant
XaIu6	Asthma monitoring by doctor
XaQig	Asthma control questionnaire
XaJ2A	Did not attend asthma clinic
XE2Nb	Asthma monitoring check done
XaR8K	Did not attend asthma review
XM1U3	Asthma clinic administration
XaY2q	Seen in school asthma clinic
Xa1hD	Acute exacerbation of asthma
XaIQ4	Change in asthma management plan
X102G	Asthmatic pulmonary eosinophilia
XaINg	Asthma limits walking on the flat

Code	Term
XaXZp	Asthma causes symptoms most nights
XaY2V	Asthma never causes night symptoms
XaXZx	Asthma limits activities most days
XaYb8	Asthma self-management plan agreed
XaYZB	Asthma self-management plan review
XaQig	ACQ - Asthma control questionnaire
XaJ2A	DNA - Did not attend asthma clinic
X1022	Intrinsic asthma with asthma attack
X101y	Extrinsic asthma with asthma attack
XaINa	Asthma never causes daytime symptoms
X1024	ASA - Aspirin-sensitive asthma triad
XaJuw	Does not have asthma management plan
XaQij	Under care of asthma specialist nurse
XaLIm	Asthma trigger - respiratory infection
Xaa7Q	No asthma trigger identified by subject
XaIIX	Asthma treatment compliance satisfactory
XE0YS	Extrinsic asthma with status asthmaticus
XaIQD	Step up change in asthma management plan
XaINd	Asthma causes daytime symptoms most days
XE0YU	Intrinsic asthma with status asthmaticus
XaIR3	Absent from work or school due to asthma
XaINf	Asthma limits walking up hills or stairs
XaRFj	Health education - asthma self-management
XaQih	Mini asthma quality of life questionnaire
XaIQE	Step down change in asthma management plan
XaIIY	Asthma treatment compliance unsatisfactory
X1024	Aspirin-sensitive asthma with nasal polyps
XE0YR	Extrinsic asthma without status asthmaticus
XaINh	Number of asthma exacerbations in past year
Xaa7B	Chronic asthma with fixed airflow obstruction
Xa544	Accidental poisoning by herbal asthma mixture
XaNKw	Royal College of Physicians asthma assessment
XaJ4W	Exception reporting: asthma quality indicators
XaXZu	Asthma limits activities 1 to 2 times per week
XaRFk	Health education - structured asthma discussion
XaXZs	Asthma limits activities 1 to 2 times per month
XaRFi	Patient has a written asthma personal action plan
XaINZ	Asthma causes night symptoms 1 to 2 times per month
XaINc	Asthma causes daytime symptoms 1 to 2 times per week
XaINb	Asthma causes daytime symptoms 1 to 2 times per month
XaXZm	Asthma causes night time symptoms 1 to 2 times per week
XaIIW	Asthma accident and emergency attendance since last visit
Xa546	Poisoning by herbal asthma mixture of undetermined intent
XaJ50	Excepted from asthma quality indicators: Informed dissent
XaXa0	Royal College of Physicians asthma assessment three questions score
XaJ4z	Excepted from asthma quality indicators: Patient unsuitable
XaX3n	Asthma review using Royal College of Physicians three questions
XaYZh	Number of days absent from school due to asthma in past 6 months

Code	Term
XaRFl	Health education - structured patient focused asthma discussion

COPD codes used fom J Quint et al's study (165)

medcode	readcode	description
794	H32..00	Emphysema
998	H3...11	Chronic obstructive airways disease
1001	H3...00	Chronic obstructive pulmonary disease
4084	663K.00	Airways obstructn irreversible
5710	H3z..00	Chronic obstructive airways disease NOS
9520	66YB.00	Chronic obstructive pulmonary disease monitoring
9876	H38..00	Severe chronic obstructive pulmonary disease
10802	H37..00	Moderate chronic obstructive pulmonary disease
10863	H36..00	Mild chronic obstructive pulmonary disease
10980	H322.00	Centrilobular emphysema
11287	66YM.00	Chronic obstructive pulmonary disease annual review
12166	H3y..00	Other specified chronic obstructive airways disease
14798	H312100	Emphysematous bronchitis
18476	66YL.11	COPD follow-up
18621	66YL.00	Chronic obstructive pulmonary disease follow-up
18792	9Oi..00	Chronic obstructive pulmonary disease monitoring admin
23492	H320z00	Chronic bullous emphysema NOS
26018	66YS.00	Chronic obstructive pulmonary disease monitoring by nurse
26306	H320.00	Chronic bullous emphysema
28755	9Oi0.00	Chronic obstructive pulmonary disease monitoring 1st letter
33450	H32z.00	Emphysema NOS
34202	9Oi1.00	Chronic obstructive pulmonary disease monitoring 2nd letter
34215	9Oi2.00	Chronic obstructive pulmonary disease monitoring 3rd letter
37247	H3z..11	Chronic obstructive pulmonary disease NOS
37371	66YD.00	Chronic obstructive pulmonary disease monitoring due
38074	9Oi4.00	Chronic obstructive pulmonary disease monitor phone invite
42258	9Oi3.00	Chronic obstructive pulmonary disease monitoring verb invite
42313	679V.00	Health education - chronic obstructive pulmonary disease
44525	H312z00	Obstructive chronic bronchitis NOS
45770	66Yg.00	Chronic obstructive pulmonary disease disturbs sleep
45771	66Yh.00	Chronic obstructive pulmonary disease does not disturb sleep
45777	8CR1.00	Chronic obstructive pulmonary disease clini management plan
45998	66YT.00	Chronic obstructive pulmonary disease monitoring by doctor
93568	H39..00	Very severe chronic obstructive pulmonary disease

Read version 3 COPD codes equivalent from above version 2 Read codes above

Version 3 Clinical Terms Version 3 codes for COPD

readcode	description
X101n	Pulmonary emphysema
H3...	Chronic obstructive pulmonary disease
Xaa7C	Chronic obstructive pulmonary disease
H3z..	Chronic obstructive airways disease NOS
XaIQT	Chronic obstructive pulmonary disease monitoring
XaIu7	Chronic obstructive pulmonary disease monitoring by nurse
XaIu8	Chronic obstructive pulmonary disease monitoring by doctor
XaEIY	Severe chronic obstructive pulmonary disease
XaN4a	Very severe chronic obstructive pulmonary disease
XaEIW	Moderate chronic obstructive pulmonary disease
XaEIV	Mild chronic obstructive pulmonary disease
XaIet	Chronic obstructive pulmonary disease annual review
H3y..	Other specified chronic obstructive airways disease
XaIes	COPD follow-up
XaIes	Chronic obstructive pulmonary disease follow-up
XaJlS	Chronic obstructive pulmonary disease monitoring admin
XaJlS	Chronic obstructive pulmonary disease monitoring administration
XE0YP	Other emphysema NOS
XaLqj	Health education - chronic obstructive pulmonary disease
H312z	Obstructive chronic bronchitis NOS
XaKv8	Chronic obstructive pulmonary disease disturbs sleep
XaKv9	Chronic obstructive pulmonary disease does not disturb sleep
XaJYf	Chronic obstructive pulmonary disease clini management plan
XaN4a	Very severe chronic obstructive pulmonary disease
XaJlT	Chronic obstructive pulmonary disease monitoring 1st letter
XaJlU	Chronic obstructive pulmonary disease monitoring 2nd letter
XaJlV	Chronic obstructive pulmonary disease monitoring 3rd letter
XaJlW	Chronic obstructive pulmonary disease monitoring verb invite
XaJlY	Chronic obstructive pulmonary disease monitoring phone invite
XaIRO	Chronic obstructive pulmonary disease monitoring due
XaIUt	Chronic obstructive pulmonary disease clini management plan given

Asthma Drug Names

- aerolin
- airomir
- albuterol
- atropine
- bambuterol
- becloforte
- beclometasone
- becodisks
- becodisk

- becotide
- betamethasone
- bitolterol
- bricanyl
- budesonide
- ciclesonide
- cobutolin
- cromogen
- cromoglicate
- cromolyn
- cyclohaler
- dexamethasone
- doxofylline
- duo-autohaler
- fenoterol
- flixotide
- flunisolide
- fluticasone
- formoterol
- indacaterol
- intal syncroner
- ipratropium
- iso-autohaler
- isoetharine
- isoprenaline
- isoproterenol
- ketotifen
- levalbuterol
- medihaler
- metaproterenol
- mometasone
- montelukast
- nedocromil
- omalizumab
- orciprenaline
- oxitropium
- oxtriphylline
- phenylephrine
- pirbuterol
- pulmicort
- reproterol
- rimiterol
- salamol
- salbutamol
- salbuvent
- seretide
- serevent
- sodium cromoglicate
- steri-neb
- symbicort
- terbutaline
- theophylline
- tinocordin
- tiotropium
- triamcinolone
- ventide
- ventolin

- zafirlukast
- zileuton

9 References

1. Catherine Foot, Chris Naylor, Candace Imison. The quality of GP diagnosis and referral [Internet]. The King's Fund; [cited 2016 Apr 24]. Available from: http://www.kingsfund.org.uk/sites/files/kf/Diagnosis%20and%20referral.pdf

2. Gkoulalas-Divanis A, Loukides G, Sun J. Publishing data from electronic health records while preserving privacy: A survey of algorithms. Journal of Biomedical Informatics. 2014 Aug;50:4–19.

3. Simon S., Jenter C, Williams D. 'Electronic Health Records: which Practices Have them and How are Clinicians Using them? '. Journal of Evaluation in Clincal Practice. 2008;14:43–7.

4. Stausberg J, Koch D, Ingenerf J, Betzler M. Comparing Paper-based with Electronic Patient Records: Lessons Learned during a Study on Diagnosis and Procedure Codes. J Am Med Inform Assoc. 2003;10(5):470–7.

5. Schade CP, Sullivan FM, de Lusignan S, Madeley J. e-Prescribing, Efficiency, Quality: Lessons from the Computerization of UK Family Practice. J Am Med Inform Assoc. 2006;13(5):470–5.

6. Jha AK, Doolan D, Grandt D, Scott T, Bates DW. The use of health information technology in seven nations. Int J Med Inform. 2008 Dec;77(12):848–54.

7. Yogeswaran P, Wright G. EHR implementation in South Africa: how do we get it right? Stud Health Technol Inform. 2010;160(Pt 1):396–400.

8. Tierney WM, Achieng M, Baker E, Bell A, Biondich P, Braitstein P, et al. Experience implementing electronic health records in three East African countries. Stud Health Technol Inform. 2010;160(Pt 1):371–5.

9. Akanbi MO, Ocheke AN, Agaba PA, Daniyam CA, Agaba EI, Okeke EN, et al. Use of Electronic Health Records in sub-Saharan Africa: Progress and challenges. J Med Trop. 2012;14(1):1–6.

10. Jawhari B, Ludwick D, Keenan L, Zakus D, Hayward R. Benefits and challenges of EMR implementations in low resource settings: a state-of-the-art review. BMC Med Inform Decis Mak. 2016;16:116.

11. World Health Organisation. WHO Atlas Report 2015 - Kenya [Internet]. [cited 2016 Sep 12]. Available from: http://www.who.int/goe/publications/atlas/2015/ken.pdf?ua=1

12. World Health Organisation. WHO Atlas Report 2015 - South Africa [Internet]. [cited 2016 Sep 12]. Available from: http://www.who.int/goe/publications/atlas/2015/zaf.pdf?ua=1

13. World Health Organisation. WHO Atlas Report 2015 - Uganda [Internet]. [cited 2016 Sep 12]. Available from: http://www.who.int/goe/publications/atlas/2015/uga.pdf?ua=1

14. World Health Organisation. WHO Atlas Report 2015 - Ghana [Internet]. [cited 2016 Sep 12]. Available from: http://www.who.int/goe/publications/atlas/2015/gha.pdf?ua=1

15. De Rosis S, Seghieri C. Basic ICT adoption and use by general practitioners: an analysis of primary care systems in 31 European countries. BMC Med Inform Decis Mak [Internet]. 2015 Aug 22 [cited 2016 Apr 4];15. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4546151/

16. Ash JS, Bates DW. Factors and Forces Affecting EHR System Adoption: Report of a 2004 ACMI Discussion. J Am Med Inform Assoc. 2005;12(1):8–12.

17. Parikh H. Overview Of EHR Systems In BRIC Nations [Internet]. [cited 2016 Apr 12]. Available from: http://www.clinicalleader.com/doc/overview-of-ehr-systems-in-bric-nations-0001

18. elizaveta.shvyndikova. EMIAS – Project of the Year [Internet]. artezio.com. 2014 [cited 2016 Apr 13]. Available from: http://www.artezio.com/pressroom/news/emias-project-year

19. Ålander T, Scandurra I. Experiences of Healthcare Professionals to the Introduction in Sweden of a Public eHealth Service: Patients' Online Access to their Electronic Health Records. Stud Health Technol Inform. 2015;216:153–7.

20. Al Alawi S, Al Dhaheri A, Al Baloushi D, Al Dhaheri M, Prinsloo EAM. Physician user satisfaction with an electronic medical records system in primary healthcare centres in Al Ain: a qualitative study. BMJ Open [Internet]. 2014 Nov 6 [cited 2016 Apr 6];4(11). Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4225459/

21. Noureldin M, Mosallam R, Hassan SZ. Quality of documentation of electronic medical information systems at primary health care units in Alexandria, Egypt. East Mediterr Health J. 2014 Feb;20(2):105–11.

22. Alsadan M, Metwally AE, Ali A, Jamal A, Khalifa M, Househ M. Health Information Technology (HIT) in Arab Countries: A Systematic Review Study on HIT Progress. Journal of Health Informatics in Developing Countries [Internet]. 2015 Jun 2 [cited 2016 Apr 1];9(2). Available from: http://www.jhidc.org/index.php/jhidc/article/view/138

23. Lejbkowicz I, Denekamp Y, Reis S, Goldenberg D. Electronic medical record systems in Israel's public hospitals. Isr Med Assoc J. 2004 Oct;6(10):583–7.

24. Dogac A, Yuksel M, Ertürkmen GL, Kabak Y, Namli T, Yıldız MH, et al. Healthcare information technology infrastructures in Turkey. Yearb Med Inform. 2014;9:228–34.

25. World Health Organisation. WHO Atlas Report 2015 - Pakistan [Internet]. [cited 2016 Sep 5]. Available from: http://www.who.int/goe/publications/atlas/2015/pak.pdf

26. World Health Organisation. WHO Atlas Report 2015 - Afghanistan [Internet]. [cited 2016 Sep 5]. Available from: http://www.who.int/goe/publications/atlas/2015/afg.pdf?ua=1

27. Sharma S. Electronic Health Record Standards For India Helpdesk | National Health Portal Of India [Internet]. [cited 2016 Apr 7]. Available from: http://www.nhp.gov.in/ehr-standards-helpdesk_ms

28. Krishnan A, Nongkynrih B, Yadav K, Singh S, Gupta V. Evaluation of computerized health management information system for primary health care in rural India. BMC Health Serv Res. 2010;10:310.

29. Radhakrishna K, Goud BR, Kasthuri A, Waghmare A, Raj T. Electronic Health Records and Information Portability: A Pilot Study in a Rural Primary Healthcare Center in India. Perspect Health Inf Manag [Internet]. 2014 Jul 1 [cited 2016 Apr 7];11(Summer). Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4142510/

30. Kulathilaka S. "eHospital - Dompe" project – the story of the transformation of a district hospital in Sri Lanka. Sri Lanka Journal of Bio-Medical Informatics [Internet].

2013 Dec 5 [cited 2016 Sep 6];4(2). Available from: http://www.sljol.info/index.php/SLJBMI/article/view/5757

31. Webster P. The rise of open-source electronic health records [Internet]. 2011 [cited 2016 Sep 6]. Available from: http://www.thelancet.com/pdfs/journals/lancet/PIIS0140-6736(11)60659-4.pdf

32. Shu T, Liu H, Goss FR, Yang W, Zhou L, Bates DW, et al. EHR adoption across China's tertiary hospitals: a cross-sectional observational study. Int J Med Inform. 2014 Feb;83(2):113–21.

33. Tanaka H. Current status of electronic health record dissemeniation in Japan [Internet]. 2007 [cited 2016 Sep 6]. Available from: http://www.med.or.jp/english/pdf/2007_05/399_404.pdf

34. Global Observatory for eHealth in Indonesia - 2013 [Internet]. [cited 2016 Apr 14]. Available from: http://www.who.int/goe/publications/atlas/2013/idn.pdf

35. Yanamadala S, Morrison D, Curtin C, McDonald K, Hernandez-Boussard T. Electronic Health Records and Quality of Care. Medicine (Baltimore) [Internet]. 2016 May 13 [cited 2016 Sep 6];95(19). Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4902473/

36. Adler-Milstein J, DesRoches CM, Furukawa MF, Worzala C, Charles D, Kralovec P, et al. More than half of US hospitals have at least a basic EHR, but stage 2 criteria remain challenging for most. Health Aff (Millwood). 2014 Sep;33(9):1664–71.

37. Jha AK, Ferris TG, Donelan K, DesRoches C, Shields A, Rosenbaum S, et al. How common are electronic health records in the United States? A summary of the evidence. Health Aff (Millwood). 2006 Dec;25(6):w496-507.

38. Chang F, Gupta N. Progress in electronic medical record adoption in Canada. Can Fam Physician. 2015 Dec;61(12):1076–84.

39. Schoen C, Osborn R, Squires D, Doty M, Rasmussen P, Pierson R, et al. A survey of primary care doctors in ten countries shows progress in use of health information technology, less in other areas. Health Aff (Millwood). 2012 Dec;31(12):2805–16.

40. Study Chronicles Electronic Health Records Successes in Mexico — MEASURE Evaluation. [cited 2016 Apr 13]; Available from: http://www.cpc.unc.edu/measure/our-work/health-information-systems/study-chronicles-electronic-health-records-successes-in-mexico

41. Hernández-Ávila JE, Palacio-Mejía LS, Lara-Esqueda A, Silvestre E, Agudelo-Botero M, Diana ML, et al. Assessing the process of designing and implementing electronic health records in a statewide public health system: the case of Colima, Mexico. J Am Med Inform Assoc. 2013 Apr;20(2):238–44.

42. Luna D, Franco M, Plaza C, Otero C, Wassermann S, Gambarte ML, et al. Accuracy of an electronic problem list from primary care providers and specialists. Stud Health Technol Inform. 2013;192:417–21.

43. World Health Organisation. WHO Atlas Report 2015 - Peru [Internet]. [cited 2016 Sep 6]. Available from: http://www.who.int/goe/publications/atlas/2015/per.pdf?ua=1

44. World Health Organisation. WHO Atlas Report 2015 - Paraguay [Internet]. [cited 2016 Sep 6]. Available from: http://www.who.int/goe/publications/atlas/2015/pry.pdf?ua=1

45. Evans J, Rogers C, Kaul S. The General Practice Morbidity Database Project Wales – A Methodology for Primary Care Data Extraction.'. Med Inform. 1997;Vol.22:No 2, 191-202.

46. de Lusignana S, van Weelb C. The use of routinely collected computer data for research in primary care: opportunities and challenges'. Family Practice. 2005 Dec 20;

47. Health Care Information Systems. In: Health Care Information Systems: A Practical Approach for Health Care Management. Jossey-Bass; 2013.

48. Treasure W. The primary care electronic health record: who's righting the software? Br J Gen Pract. 2011 Feb 1;61(583):152–4.

49. TPP Plc. SystmOne Modules - systmOne GP [Internet]. TPP. [cited 2016 Apr 19]. Available from: http://www.tpp-uk.com/products/systmone/modules/

50. British Medical Journal. GP Systems of Choice (GPSoC) Questions & Answers [Internet]. Health Analytics: Clinical Software Systems & Services. [cited 2016 Apr 19]. Available from: http://healthanalytics.bmj.com/customers/gpsoc/#faq1

51. The Health and Social Care Information Centre (HSCIC). GPSoC — Health and Social Care Information Centre [Internet]. [cited 2016 Apr 19]. Available from: http://systems.hscic.gov.uk/gpsoc/about

52. Waterson P. Health information technology and sociotechnical systems: a progress report on recent developments within the UK National Health Service (NHS). Appl Ergon. 2014 Mar;45(2):150–61.

53. Carlisle George. Overview of national legislation on EHR in the UK (England) [Internet]. 2014 [cited 2016 Apr 15]. Available from: http://ec.europa.eu/health/ehealth/docs/laws_united_kingdom_en.pdf

54. TPP to become second biggest GP supplier [Internet]. TPP. 2011 [cited 2016 Mar 3]. Available from: http://www.tpp-uk.com/latest-news-stories/tpp-to-become-second-biggest-gp-supplier/

55. TPP Plc. SystmOnline [Internet]. TPP. [cited 2016 Apr 20]. Available from: http://www.tpp-uk.com/products/systmonline/

56. Bernard Ginns. Emis Group at the vanguard of digital revolution in NHS [Internet]. 2014 [cited 2016 Apr 19]. Available from: http://www.yorkshirepost.co.uk/news/emis-group-at-the-vanguard-of-digital-revolution-in-nhs-1-6390769

57. Emis Group: Our History [Internet]. [cited 2016 Apr 17]. Available from: http://www.emisgroupplc.com/who-we-are/our-history/

58. Emis Health. EMIS to withdraw legacy systems in England following changes to GP computing requirements [Internet]. 2014 [cited 2016 Apr 19]. Available from: http://www.emishealth.com/news-events/news/emis-to-withdraw-legacy-systems-in-england-following-changes-to-gp-computing-requirements/

59. J Raftery, P Roderick, A Stevens. Potential use of routine databases in health technology assessment - Appendices [Internet]. [cited 2016 Apr 19]. Available from: http://www.journalslibrary.nihr.ac.uk/__data/assets/pdf_file/0009/59544/Appendices-hta9200.pdf

60. INPS. GP practices | INPS [Internet]. [cited 2016 Apr 19]. Available from: http://www.inps.co.uk/vision/gp-practices

61. CSC Finalizes iSOFT Acquisition [Internet]. [cited 2016 Apr 20]. Available from: http://www.csc.com/newsroom/press_releases/68289-csc_finalizes_isoft_acquisition

62. Computer Sciences Corporation. iSOFT - Healthcare Systems in the UK [Internet]. Computer Sciences Corporation. [cited 2016 Apr 19]. Available from: http://www.csc.com/uk/ahp/100585-healthcare_in_the_uk

63. Microtest Heal. Evolution [Internet]. Microtest Health Solutions. [cited 2016 Apr 19]. Available from: http://www.microtest.co.uk/evolution/

64. Crosscare Primary - Advanced Health & Care [Internet]. [cited 2016 Apr 20]. Available from: http://www.advancedcomputersoftware.com/ahc/products/crosscare-practice-management-system.php

65. BMA - GP contract IT requirements [Internet]. [cited 2016 Sep 7]. Available from: https://www.bma.org.uk/advice/employment/contracts/general-practice-funding/gp-contract-2015-2016-england/gp-contract-it-requirements

66. Luchenski SA, Reed JE, Marston C, Papoutsi C, Majeed A, Bell D. Patient and public views on electronic health records and their uses in the United kingdom: cross-sectional survey. J Med Internet Res. 2013;15(8):e160.

67. World Health Organization. International Classification of Diseases (ICD) [Internet]. WHO. [cited 2014 Jun 26]. Available from: http://www.who.int/classifications/icd/en/

68. Rollason K, Khunti K, de Lusignan S. Variation in the recording of diabetes diagnostic data in primary care computer systems: Implications for the quality of care. Informatics in Primary Care , British Computer Society. 2009;17:17:113-119.

69. Benson T. The history of the Read codes: the inaugural James Read Memorial Lecture 2011. Informatics in Primary Care. 2011 Jun 1;19(3):173–82.

70. Business rules v17.0 - NHS Primary Care Commissioning [Internet]. [cited 2011 Mar 2]. Available from: http://www.pcc.nhs.uk/business-rules-v17-0

71. Gray J, Orr D, Majeed A. 'Information in practice Use of Read codes in diabetes management in a south London primary care group: implications for establishing disease registers.' BMJ. 2003 May 24;VOLUME 326.

72. Wood L, Coulson R. Revitalizing the General Practice Research Database: Plans, Challenges and Opportunities. Pharmacoepidemiology and drug safety. 2001;10; 379-383.

73. Quality and Outcomes Framework - 2009-10 [Internet]. United Kingdom: Health and Social Care Information Centre; 2010 Jun [cited 2015 Sep 28]. Available from: http://www.hscic.gov.uk/catalogue/PUB04431

74. The NHS Information Centre, Prescr. Quality and Outcomes Framework Achievement Data 2009/10 [Internet]. 2010 [cited 2016 Apr 24]. Available from: http://www.hscic.gov.uk/catalogue/PUB04431/qof-09-10-rep.pdf

75. (mr) Web Master UK. Quality and Outcomes Framework (QOF) - 2015-16 [Internet]. 2016 [cited 2017 Jan 3]. Available from: http://www.content.digital.nhs.uk/catalogue/PUB22266

76. Sean Canderhill. Health Economics: Database Studies [Internet]. RTI Health Solutions; 2016 [cited 2016 Apr 15]. Available from: https://www.rtihs.org/services/health-economics/health-economics-database-studies

77. Hippisley-Cox J, Stables D, Pringle M. QRESEARCH: a new general practice database for research'. Informatics in Primary Care 2004; 2004 PHCSG, British Computer Society.

78. Barson R. National award for transforming data into knowledge [Internet]. [cited 2016 Mar 3]. Available from: https://www.leeds.ac.uk/news/article/3628/national_award_for_transforming_data_into_knowledge

79. Bourke A, Dattani H, Robinson M. Feasibility study and methodology to create a quality-evaluated database of primary care data. Inform Prim Care. 2004;12(3):171–7.

80. Cegedim Strategic Data Medical Research UK. The Health Improvement Network [Internet]. [cited 2014 Apr 6]. Available from: http://www.thin-uk.com/

81. Carr MJ, Ashcroft DM, Kontopantelis E, Awenat Y, Cooper J, Chew-Graham C, et al. The epidemiology of self-harm in a UK-wide primary care patient cohort, 2001-2013. BMC Psychiatry. 2016;16(1):53.

82. Grimaldi-Bensouda L, Klungel O, Kurz X, de Groot MCH, Maciel Afonso AS, de Bruin ML, et al. Calcium channel blockers and cancer: a risk analysis using the UK Clinical Practice Research Datalink (CPRD). BMJ Open. 2016;6(1):e009147.

83. Gallagher AM, Williams T, Leufkens HGM, de Vries F. The Impact of the Choice of Data Source in Record Linkage Studies Estimating Mortality in Venous Thromboembolism. PLoS ONE. 2016;11(2):e0148349.

84. MHRA. Clinical Practice Research Datalink - CPRD [Internet]. [cited 2016 Apr 23]. Available from: https://www.cprd.com/home/#nogo

85. QResearch - What Is QResearch [Internet]. [cited 2011 Aug 19]. Available from: http://www.qresearch.org/SitePages/What%20Is%20QResearch.aspx

86. Q Research. Researchers guide to QRESEARCH version 1.5 [Internet]. [cited 2016 Sep 13]. Available from: http://www.qresearch.org/Public_Documents/Researchers%20guide%20to%20QRESEARCH%20version%201.5.pdf

87. TPP launches ResearchOne [Internet]. TPP. 2012 [cited 2016 Apr 23]. Available from: http://www.tpp-uk.com/latest-news-stories/tpp-launches-researchone/

88. Crossfield S. TPP 5 Million Records Support Research in 2013 | ResearchOne [Internet]. [cited 2016 Apr 23]. Available from: http://www.researchone.org/400-organisations-join-in-2013/

89. IMS Health : Data collection [Internet]. [cited 2016 Mar 8]. Available from: http://www.epic-uk.org/our-data/data-content.shtml

90. Geochemical baselines | G-BASE | Our research | British Geological Survey (BGS) [Internet]. [cited 2016 Mar 8]. Available from: http://www.bgs.ac.uk/gbase/home.html

91. Lewis S, Sharma A, Szatkowski L. Insights into social disparities in smoking prevalence using Mosaic, a novel measure of socioeconomic status: an analysis using a large primary care dataset. BMC Public Health 2010. 2010 Dec 7;(10:755).

92. Cegedim Strategic Data Medical Research UK. THIN Data Guide for Researchers - Version 2.0. 2010.

93. CPRD Record Linkage [Internet]. [cited 2016 Mar 8]. Available from: https://www.cprd.com/recordLinkage/

94. Availability_and_use_of_UK_based_ethnicity_data_for_health_res_1.pdf [Internet]. [cited 2016 Mar 8]. Available from: http://eprints.ncrm.ac.uk/3040/1/Mathur-_Availability_and_use_of_UK_based_ethnicity_data_for_health_res_1.pdf

95. TPP SystmOne Research Database - ResearchOne-Database-System-Summary_V2.pdf [Internet]. [cited 2016 Mar 8]. Available from: http://www.researchone.org/wp-content/uploads/2013/09/ResearchOne-Database-System-Summary_V2.pdf

96. Weng C, Appelbaum P, Hripcsak G, Kronish I, Busacca L, Davidson KW, et al. Using EHRs to integrate research with patient care: promises and challenges. J Am Med Inform Assoc. 2012;19(5):684–7.

97. Menachemi N, Collum TH. Benefits and drawbacks of electronic health record systems. Risk Manag Healthc Policy. 2011 May 11;4:47–55.

98. van Staa T-P, Dyson L, McCann G, Padmanabhan S, Belatri R, Goldacre B, et al. The opportunities and challenges of pragmatic point-of-care randomised trials using routinely collected electronic records: evaluations of two exemplar trials. Health Technol Assess. 2014 Jul;18(43):1–146.

99. Improving Patient Outcomes and Diagnostics with EHRs | Providers & Professionals | HealthIT.gov [Internet]. [cited 2016 Apr 20]. Available from: https://www.healthit.gov/providers-professionals/improved-diagnostics-patient-outcomes

100. Vawdrey DK, Weng C, Herion D, Cimino JJ. Enhancing Electronic Health Records to Support Clinical Research. AMIA Jt Summits Transl Sci Proc. 2014 Apr 7;2014:102–8.

101. Britton J, Pavord I, Richards K, Wisniewski A, Knox A, Lewis S, et al. Dietary magnesium, lung function, wheezing, and airway hyperreactivity in a random adult population sample. Lancet. 1994 Aug 6;344(8919):357–62.

102. Lewis SA, Weiss ST, Britton JR. Airway responsiveness and peak flow variability in the diagnosis of asthma for epidemiological studies. Eur Respir J. 2001 Dec;18(6):921–7.

103. McKeever TM, Scrivener S, Broadfield E, Jones Z, Britton J, Lewis SA. Prospective Study of Diet and Decline in Lung Function in a General Population. American Journal of Respiratory and Critical Care Medicine. 2002 May;165(9):1299–303.

104. Britton JR, Pavord ID, Richards KA, Knox AJ, Wisniewski AF, Lewis SA, et al. Dietary antioxidant vitamin intake and lung function in the general population. Am J Respir Crit Care Med. 1995 May 1;151(5):1383–7.

105. Britton J, Pavord I, Richards K, Knox A, Wisniewski A, Wahedna I, et al. Factors influencing the occurrence of airway hyperreactivity in the general population: the importance of atopy and airway calibre. Eur Respir J. 1994 May;7(5):881–7.

106. Britton J, Pavord I, Richards K, Knox A, Wisniewski A, Weiss S, et al. Dietary sodium intake and the risk of airway hyperreactivity in a random adult population. Thorax. 1994 Sep;49(9):875–80.

107. McKeever TM, Scrivener S, Broadfield E, Jones Z, Britton J, Lewis SA. Prospective study of diet and decline in lung function in a general population. Am J Respir Crit Care Med. 2002 May 1;165(9):1299–303.

108. Broadfield E, McKeever TM, Scrivener S, Venn A, Lewis SA, Britton J. Increase in the prevalence of allergen skin sensitization in successive birth cohorts. Journal of Allergy and Clinical Immunology. 2002 Jun;109(6):969–74.

109. Database Data Warehousing Guide [Internet]. [cited 2016 Feb 12]. Available from: https://docs.oracle.com/cd/B19306_01/server.102/b14223/extract.htm#CHDGGFEA

110. Powell J, Buchan I. Electronic Health Records Should Support Clinical Research. Journal of Medical Internet Research Med Internet Res [Internet]. 2005 Mar 14;2005 Jan-Mar; 7(1). Available from: http//jmir.org

111. Liaw S-T, Taggart J, Yu H, de Lusignan S. Data extraction from electronic health records - existing tools may be unreliable and potentially unsafe. Aust Fam Physician. 2013 Nov;42(11):820–3.

112. Wang Y, Steinhubl SR, Defilippi C, Ng K, Ebadollahi S, Stewart WF, et al. Prescription Extraction from Clinical Notes: Towards Automating EMR Medication Reconciliation. AMIA Jt Summits Transl Sci Proc. 2015;2015:188–93.

113. Tiro JA, Sanders JM, Shay LA, Murphy CC, A Hamann H, Bartholomew LK, et al. Validation of self-reported post-treatment mammography surveillance among breast cancer survivors by electronic medical record extraction method. Breast Cancer Res Treat. 2015 Jun;151(2):427–34.

114. Elkhenini HF, Davis KJ, Stein ND, New JP, Delderfield MR, Gibson M, et al. Using an electronic medical record (EMR) to conduct clinical trials: Salford Lung Study feasibility. BMC Med Inform Decis Mak. 2015;15:8.

115. Holden RJ. What stands in the way of technology-mediated patient safety improvements?: a study of facilitators and barriers to physicians' use of electronic health records. J Patient Saf. 2011 Dec;7(4):193–203.

116. Harris SB, Glazier RH, Tompkins JW, Wilton AS, Chevendra V, Stewart MA, et al. Investigating concordance in diabetes diagnosis between primary care charts (electronic medical records) and health administrative data: a retrospective cohort study. BMC Health Serv Res. 2010;10:347.

117. Bobo WV, Pathak J, Kremers HM, Yawn BP, Brue SM, Stoppel CJ, et al. An electronic health record driven algorithm to identify incident antidepressant medication users. J Am Med Inform Assoc. 2014 Oct;21(5):785–91.

118. Lobach DF, Detmer DE. Research challenges for electronic health records. Am J Prev Med. 2007 May;32(5 Suppl):S104-111.

119. Wasserman RC. Electronic medical records (EMRs), epidemiology, and epistemology: reflections on EMRs and future pediatric clinical research. Acad Pediatr. 2011 Aug;11(4):280–7.

120. Chung S-C, Sundström J, Gale CP, James S, Deanfield J, Wallentin L, et al. Comparison of hospital variation in acute myocardial infarction care and outcome between Sweden and United Kingdom: population based cohort study using nationwide clinical registries. BMJ. 2015;351:h3913.

121. Adamusiak T, Shimoyama N, Shimoyama M. Next generation phenotyping using the unified medical language system. JMIR Med Inform. 2014;2(1):e5.

122. Tu K, Widdifield J, Young J, Oud W, Ivers NM, Butt DA, et al. Are family physicians comprehensively using electronic medical records such that the data can be used for secondary purposes? A Canadian perspective. BMC Med Inform Decis Mak. 2015;15:67.

123. Kadra G, Stewart R, Shetty H, Jackson RG, Greenwood MA, Roberts A, et al. Extracting antipsychotic polypharmacy data from electronic health records: developing and evaluating a novel process. BMC Psychiatry. 2015;15:166.

124. Boyer C, Dolamic L, Grabar N. Automated Detection of Health Websites' HONcode Conformity: Can N-gram Tokenization Replace Stemming? Stud Health Technol Inform. 2015;216:1064.

125. Doan S, Maehara CK, Chaparro JD, Lu S, Liu R, Graham A, et al. Building a Natural Language Processing Tool to Identify Patients with High Clinical Suspicion for Kawasaki Disease from Emergency Department Notes. Acad Emerg Med. 2016 Jan 30;

126. Cai T, Giannopoulos AA, Yu S, Kelil T, Ripley B, Kumamaru KK, et al. Natural Language Processing Technologies in Radiology Research and Clinical Applications. Radiographics. 2016 Feb;36(1):176–91.

127. Mishra NK, Son RY, Arnzen JJ. Towards Automatic Diabetes Case Detection and ABCS Protocol Compliance Assessment. Clin Med Res. 2012 Aug;10(3):106–21.

128. Kutsia T, Marin M. Regular expression order-sorted unification and matching. J Symb Comput. 2015 Mar;67:42–67.

129. Redd D, Kuang J, Mohanty A, Bray BE, Zeng-Treitler Q. Regular Expression-Based Learning for METs Value Extraction. AMIA Jt Summits Transl Sci Proc. 2016;2016:213–20.

130. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J Am Med Inform Assoc. 2013 Jan 1;20(1):144–51.

131. Karystianis G, Sheppard T, Dixon WG, Nenadic G. Modelling and extraction of variability in free-text medication prescriptions from an anonymised primary care electronic medical record research database. BMC Med Inform Decis Mak. 2016;16(1):18.

132. Evans DA, Brownlow ND, Hersh WR, Campbell EM. Automating concept identification in the electronic medical record: an experiment in extracting dosage information. Proc AMIA Annu Fall Symp. 1996;388–92.

133. Gold S, Elhadad N, Zhu X, Cimino JJ, Hripcsak G. Extracting structured medication event information from discharge summaries. AMIA Annu Symp Proc. 2008 Nov 6;237–41.

134. Voorham J, Denig P. Computerized extraction of information on the quality of diabetes care from free text in electronic patient records of general practitioners. J Am Med Inform Assoc. 2007 Jun;14(3):349–54.

135. Turchin A, Kolatkar NS, Grant RW, Makhni EC, Pendergrass ML, Einbinder JS. Using Regular Expressions to Abstract Blood Pressure and Treatment Intensification Information from the Text of Physician Notes. J Am Med Inform Assoc. 2006;13(6):691–5.

136. Xu H, Stenner SP, Doan S, Johnson KB, Waitman LR, Denny JC. MedEx: a medication information extraction system for clinical narratives. J Am Med Inform Assoc. 2010 Feb;17(1):19–24.

137. Hinchcliff M, Just E, Podlusky S, Varga J, Chang RW, Kibbe WA. Text data extraction for a prospective, research-focused data mart: implementation and validation. BMC Med Inform Decis Mak. 2012 Sep 13;12:106.

138. Regular Expression Syntax [Internet]. 2010 [cited 2016 Nov 8]. Available from: https://msdn.microsoft.com/en-us/library/ae5bf541(v=vs.100).aspx

139. Moskow JM, Cook N, Champion-Lippmann C, Amofah SA, Garcia AS. Identifying opportunities in EHR to improve the quality of antibiotic allergy data. J Am Med Inform Assoc. 2015 Nov 9;

140. Cochrane NJK. Programmatic extraction of information from unstructured clinical data and the assessment of potential impacts on epidemiological research [Internet]. 2015 [cited 2016 Nov 28]. Available from: http://eprints.nottingham.ac.uk/30582/

141. Health and Social Care Information Centre. Technology Reference data Update Distribution (TRUD) 15.0.0 Release Pack [Internet]. 2014. Available from: https://isd.hscic.gov.uk/

142. gnu.org [Internet]. [cited 2017 Jul 27]. Available from: https://www.gnu.org/licenses/gpl-3.0.en.html

143. Reeves D, Springate DA, Ashcroft DM, Ryan R, Doran T, Morris R, et al. Can analyses of electronic patient records be independently and externally validated? The effect of statins on the mortality of patients with ischaemic heart disease: a cohort study with nested case–control analysis. BMJ Open. 2014 Apr 1;4(4):e004952.

144. Springate DA, Ashcroft DM, Kontopantelis E, Doran T, Ryan R, Reeves D. Can analyses of electronic patient records be independently and externally validated? Study 2--the effect of β-adrenoceptor blocker therapy on cancer survival: a retrospective cohort study. BMJ Open. 2015;5(4):e007299.

145. Taggar JS, Coleman T, Lewis S, Szatkowski L. The impact of the Quality and Outcomes Framework (QOF) on the recording of smoking targets in primary care medical records: cross-sectional analyses from The Health Improvement Network (THIN) database. BMC Public Health. 2012;12:329.

146. Tate AR, Martin AGR, Ali A, Cassell JA. Using free text information to explore how and when GPs code a diagnosis of ovarian cancer: an observational study using primary care records of patients with ovarian cancer. BMJ Open. 2011 Jan 1;1(1):e000025.

147. The Area Based Analysis Unit, Office for National Statistics. Understanding patterns of deprivation [Internet]. Office for National Statistics. 2010 [cited 2016 May 8]. Available from: http://web.ons.gov.uk/ons/publications/index.html

148. Rose D. Social Research Update 9: Official Social Classifications in the UK [Internet]. [cited 2016 May 8]. Available from: http://sru.soc.surrey.ac.uk/SRU9.html

149. Ford E, Nicholson A, Koeling R, Tate A, Carroll J, Axelrod L, et al. Optimising the use of electronic health records to estimate the incidence of rheumatoid arthritis in primary care: what information is hidden in free text? BMC Med Res Methodol. 2013;13:105.

150. Wang Z, Shah AD, Tate AR, Denaxas S, Shawe-Taylor J, Hemingway H. Extracting diagnoses and investigation results from unstructured text in electronic health records by semi-supervised machine learning. PLoS ONE. 2012;7(1):e30412.

151. Valkhoff VE, Coloma PM, Masclee GMC, Gini R, Innocenti F, Lapi F, et al. Validation study in four health-care databases: upper gastrointestinal bleeding misclassification affects precision but not magnitude of drug-related upper gastrointestinal bleeding risk. J Clin Epidemiol. 2014 Aug;67(8):921–31.

152. Gundlapalli AV, Redd D, Gibson BS, Carter M, Korhonen C, Nebeker J, et al. Maximizing clinical cohort size using free text queries. Comput Biol Med. 2015 May;60:1–7.

153. Taggart J, Liaw S-T, Yu H. Structured data quality reports to improve EHR data quality. Int J Med Inform. 2015 Dec;84(12):1094–8.

154. Fisher ES, McClellan MB, Safran DG. Building the path to accountable care. N Engl J Med. 2011 Dec 29;365(26):2445–7.

155. Kontopantelis E, Buchan I, Reeves D, Checkland K, Doran T. Relationship between quality of care and choice of clinical computing system: retrospective analysis of family practice performance under the UK's quality and outcomes framework. BMJ Open [Internet]. 2013 Aug 2 [cited 2016 Jul 20];3(8). Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3733310/

156. Kanger C, Brown L, Mukherjee S, Xin H, Diana ML, Khurshid A. Evaluating the Reliability of EHR-Generated Clinical Outcomes Reports: A Case Study. EGEMS (Wash DC) [Internet]. 2014 Oct 23 [cited 2016 Mar 11];2(3). Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4371440/

157. Turchin A, Sawarkar A, Dementieva YA, Breydo E, Ramelson H. Effect of EHR user interface changes on internal prescription discrepancies. Appl Clin Inform. 2014;5(3):708–20.

158. Holt TA, Gunnarsson CL, Cload PA, Ross SD. Identification of undiagnosed diabetes and quality of diabetes care in the United States: cross-sectional study of 11.5 million primary care electronic records. CMAJ Open. 2014 Oct 1;2(4):E248–55.

159. Weiskopf NG, Rusanov A, Weng C. Sick patients have more data: the non-random completeness of electronic health records. AMIA Annu Symp Proc. 2013;2013:1472–7.

160. West SL, Strom BL, Freundlich B, Normand E, Koch G, Savitz DA. Completeness of prescription recording in outpatient medical records from a health maintenance organization. J Clin Epidemiol. 1994 Feb;47(2):165–71.

161. Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. J Am Med Inform Assoc. 2013 Jan 1;20(1):117–21.

162. Chau N, Chau K, Mayet A, Baumann M, Legleye S, Falissard B. Self-reporting and measurement of body mass index in adolescents: refusals and validity, and the possible role of socioeconomic and health-related factors. BMC Public Health. 2013 Sep 8;13:815.

163. Dhalwani NN, Tata LJ, Coleman T, Fleming KM, Szatkowski L. Completeness of maternal smoking status recording during pregnancy in United Kingdom primary care data. PLoS ONE. 2013;8(9):e72218.

164. Mukherjee M, Stoddart A, Gupta RP, Nwaru BI, Farr A, Heaven M, et al. The epidemiology, healthcare and societal burden and costs of asthma in the UK and its member nations: analyses of standalone and linked national databases. BMC Med. 2016;14(1):113.

165. Quint JK, Müllerova H, DiSantostefano RL, Forbes H, Eaton S, Hurst JR, et al. Validation of chronic obstructive pulmonary disease recording in the Clinical Practice Research Datalink (CPRD-GOLD). BMJ Open. 2014;4(7):e005540.

166. National Institute Clinical Excellence. Chronic obstructive pulmonary disease in over 16s: diagnosis and management | 1-guidance | Guidance and guidelines | NICE [Internet]. [cited 2016 Jun 9]. Available from: https://www.nice.org.uk/guidance/CG101/chapter/1-Guidance#diagnosing-copd

167. McKeever T, Harrison TW, Hubbard R, Shaw D. Inhaled corticosteroids and the risk of pneumonia in people with asthma: a case-control study. Chest. 2013 Dec;144(6):1788–94.

168. Furu K, Karlstad Ø, Skurtveit S, Håberg SE, Nafstad P, London SJ, et al. High validity of mother-reported use of antiasthmatics among children: a comparison with a population-based prescription database. J Clin Epidemiol. 2011 Aug;64(8):878–84.

169. Lyons N, Rankin S, Sarangarm P, Washington C, Weiss SJ, Ernst AA. Disparity in Patients' Self-Reported and Charted Medication Allergy Information. South Med J. 2015 Jun;108(6):332–6.

170. Fernando B, Morrison Z, Kalra D, Cresswell K, Sheikh A. Approaches to Recording Drug Allergies in Electronic Health Records: Qualitative Study. PLoS One [Internet]. 2014 Apr 16 [cited 2016 Dec 13];9(4). Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3989180/

171. Xi N, Wallace R, Agarwal G, Chan D, Gershon A, Gupta S. Identifying patients with asthma in primary care electronic medical record systems. Can Fam Physician. 2015 Oct;61(10):e474–83.

172. Rothnie KJ, Müllerová H, Hurst JR, Smeeth L, Davis K, Thomas SL, et al. Validation of the Recording of Acute Exacerbations of COPD in UK Primary Care Electronic Healthcare Records. PLoS ONE. 2016;11(3):e0151357.

173. Powell HA, Iyen-Omofoman B, Baldwin DR, Hubbard RB, Tata LJ. Chronic Obstructive Pulmonary Disease and Risk of Lung Cancer: The Importance of Smoking and Timing of Diagnosis. Journal of Thoracic Oncology. 2013 Jan;8(1):6–11.

174. Weiskopf NG, Hripcsak G, Swaminathan S, Weng C. Defining and measuring completeness of electronic health records for secondary use. J Biomed Inform. 2013 Oct;46(5):830–6.

175. Campbell SM, Reeves D, Kontopantelis E, Sibbald B, Roland M. Effects of pay for performance on the quality of primary care in England. N Engl J Med. 2009 Jul 23;361(4):368–78.

176. Campbell S, Reeves D, Kontopantelis E, Middleton E, Sibbald B, Roland M. Quality of primary care in England with the introduction of pay for performance. N Engl J Med. 2007 Jul 12;357(2):181–90.

177. Smith CJP, Gribbin J, Challen KB, Hubbard RB. The impact of the 2004 NICE guideline and 2003 General Medical Services contract on COPD in primary care in the UK. QJM. 2008 Feb;101(2):145–53.

178. Maisey S, Steel N, Marsh R, Gillam S, Fleetcroft R, Howe A. Effects of payment for performance in primary care: qualitative interview study. J Health Serv Res Policy. 2008 Jul;13(3):133–9.

179. Fleetcroft R, Noble M, Martin A, Coombes E, Ford J, Steel N. Emergency hospital admissions for asthma and access to primary care: cross-sectional analysis. Br J Gen Pract. 2016 Jun 20;

180. Department of Health. Quality and Outcomes Framework (QOF) [Internet]. 2009 [cited 2009 May 12]. Available from: http://www.dh.gov.uk/en/Healthcare/Primarycare/Primarycarecontracting/QOF/index.htm

181. Watson AJ, O'Rourke J, Jethwani K, Cami A, Stern TA, Kvedar JC, et al. Linking electronic health record-extracted psychosocial data in real-time to risk of readmission for heart failure. Psychosomatics. 2011;52(4):319–27.

182. Garvin JH, DuVall SL, South BR, Bray BE, Bolton D, Heavirland J, et al. Automated extraction of ejection fraction for quality measurement using regular expressions in

Unstructured Information Management Architecture (UIMA) for heart failure. J Am Med Inform Assoc. 2012;19(5):859–66.

183. Shah AD, Martinez C, Hemingway H. The freetext matching algorithm: a computer program to extract diagnoses and causes of death from unstructured text in electronic health records. BMC Med Inform Decis Mak. 2012;12:88.

184. Ross MK, Wei W, Ohno-Machado L. "Big data" and the electronic health record. Yearb Med Inform. 2014;9:97–104.

185. Murdoch TB, Detsky AS. The inevitable application of big data to health care. JAMA. 2013 Apr 3;309(13):1351–2.

186. PRIMIS quality improvement tools [Internet]. 2016 [cited 2016 Dec 20]. Available from: http://www.nottingham.ac.uk/primis/tools-audits/index.aspx