HERALD (Health Economics using Routine Anonymised Linked Data) Protocol Paper

Husain M.J[1*], Macey S[1], Brophy S[1], Pinder L[1], Atkinson MD[1], Cooksey R[1], Phillips CJ[2], Siebert S[1.]

1. College of Medicine, Swansea University, Wales. SA2 8PP
2. College of Human and Health Sciences, Swansea University, Wales. SA2 8PP

Email addresses

MJH: m.j.husain@swansea.ac.uk

SM: s.m.macey@swansea.ac.uk.

SB: s.brophy@swansea.ac.uk

LP: l.pinder@swansea.ac.uk

MDA: m.atkinson@swansea.ac.uk

RC: r.cooksey@swansea.ac.uk

CJP: c.j.phillips@swansea.ac.uk

SS: s.siebert@swansea.ac.uk

*** Corresponding author:** M. Jami Husain at m.j.husain@swansea.ac.uk Tel: 01792 606621. College of Medicine, Swansea University.

Abstract

Background

Health economic analysis tends to rely on patient derived questionnaire data, routine datasets, and resource and outcomes data from experimental randomised control trials and other clinical studies. However, these data sources are generally used as stand-alone datasets. Herein, we outline the potential implications of linking these datasets to give one single joined up data-resource for health economic analysis.

Method

The linkage of individual level data from questionnaires with routinely-captured health care data allows the entire patient journey to be mapped both prospectively and retrospectively. We illustrate this with examples from an Ankylosing Spondylitis cohort by linking patient reported study dataset with the routinely collected general practitioner data; inpatient and outpatient datasets; and Accident and Emergency department data in the United Kingdom.

Results

The potential implications of linking routine data with study data are: (1) The patient-reported recall data can be cross-checked with actual events of health care and social care interventions at the personal level. (2) The patient pathway through the healthcare system can be tracked, both retrospectively and prospectively. This means the questionnaire burden on patients can be reduced and objective measures of health service use can be captured over a time period that is not feasible using patient recall, providing a more objective estimate of the true costs of chronic conditions and any interventions. (3) A retrospective examination of patients' histories can allow an estimation of costs during the pre-diagnosis period. (4) A retrospective/prospective analysis of patients' healthcare histories can offer an assessment of health service usage and recommendations for improving patient care pathways. (5) The

linked data enables profiling and stratification of patients relating to disease manifestation, lifestyles, co-morbidities, and associated costs.

Conclusion

The use of this combined model provides an objective estimate of the cost and potential burden of disease to the health care system, society and the individual at each stage of disease over a prolonged period of time. The ability to link data not only enhances the accuracy and reliability of the cost estimates but additionally acts as a unique validation tool, allowing potential limitations/strengths of each source of data collection to be identified.

Introduction

Health service research have tended to rely on data generated from randomised controlled trials (RCT), observational studies based on patient derived questionnaire data, and routinely assembled data abstracted from the primary and secondary care patient record - popularly known as routine data [1]. Data generated from these three routes are predominantly used as stand-alone datasets, and alongside non-health data such as demographic and geographical data [1, 2]. Intuitively, the health sector analyses would be enriched if the patient-level data generated from these different sources could be linked. This paper discusses the potential methodological advantages in the conduct of health economics analyses using patient-derived questionnaire data, linked with routinely collected information and secondary care clinical datasets.

Methods

SAIL databank

In order to realise the potential of electronically-held routinely collected information to conduct and support health-related research, the Health Information Research Unit (HIRU) at the College of Medicine at Swansea University, as part of the Welsh Assembly Government's commitment to the UK Clinical Research Collaboration (UKCRC), has set up the Secure Anonymised Information Linkage (SAIL) databank [3, 4]. The SAIL databank brings together and links a wide range of person-based data. SAIL utilises a split-file approach to anonymisation to overcome issues of confidentiality and disclosure in health-related data warehousing by creating personal-level unique and encrypted identifiers for merging information from various sources [3, 4]. The range of complementary sets of data includes clinical data from rheumatologists, existing routinely collected datasets such as the General

Practice (GP) records, out-patients clinical data, inpatient episodes, accident and emergency department, pathology data and social services databases.

Data linkage

Data collected through patient questionnaires can be linked to other routinely collected datasets using the SAIL system. HIRU uses the MACRAL (Matching Algorithm for Consistent Results in Anonymised Linkage) algorithm to create encrypted Anonymised Linking Field (ALF) for each individual [3]. The ALFs are mainly created based on the patient's NHS number; and if the NHS number is absent in a dataset, a mixture of other identifying variables like forename, surname, gender, postcode of residence, and date of birth are used for probabilistic matching, while maintaining complete anonymity for the end users [3]. This linkage allows us to follow the patient pathway through the NHS system both retrospectively and prospectively.

Results

Potential benefits of using linked data

1. Validation of patient-reported recall data

The use of linked routine data allows cross-checking of patient-reported recall data with actual events of health care interventions at the personal level. The inherent limitations (or strengths) of the data quality pertaining to the survey questionnaires under the recall method can be flagged; and an assessment of the generalisability of the patient-reported data can be made. On the other hand, data obtained from routinely collected data systems often require careful interpretation with respect to their quality, validity, timelines, bias, confounding and statistical stability [1]. With the triangulation of datasets in the SAIL system, the validity and reliability of single datasets can be assessed [5].

2. Retrospective and prospective tracking of patient pathways

With data spanning multiple years and the ability to link records across several datasets, it is possible using SAIL to track the healthcare utilisation history of patients in receipt of some form of treatment for a given condition across multiple healthcare sectors both before and after their index/reference healthcare event. Therefore, SAIL data linkage system allows tracking of the patient pathways, both retrospectively and prospectively. Linkage with GP data system provides information about patients' primary care events going back many years including; previous diagnoses, referrals, presenting symptoms, investigation results and previous medications. This dataset can also be used to follow the patient at every visit to the GP and therefore record the development of associated conditions and use of co-medications. Linkage with inpatient data will record all hospital visits, surgery and hospital treatment. Linkage with the mortality datasets will ensure the dataset remains relevant and can examine survival of included patients. Linkage with Accident and Emergency (A&E) datasets will give information on emergency visits.

Figure 1 shows the linkable data sets available in the SAIL databank, and shows the timelines of data availability for each of them. This growing databank already holds over a billion anonymised records from the listed databases and these can be anonymously linked at the individual record level, allowing tracking of patients' health related history retrospectively. For instance, data from the GP primary care, NHS administrative register, breast and cervical cancer screening extend back two decades.

3. Cost of illness in pre-diagnosis stage

Cost of illness studies are typically subject to a degree of scrutiny with regards to the sources and methods of estimating the quantities and prices, the specification of study perspective, and the identification of the timeframe to which the costs apply [6, 7, 8]. The use of linked

data system enhances the precision of the healthcare use information and the timelines within which the costs incurred. Therefore, the evaluation using the HERALD (Health Economics using Routine Anonymised Linked Data) methodology will help provide an objective estimate of the cost and burden of diseases to the United Kingdom NHS, society and the individual at each stage of disease over a prolonged period of time.

In many conditions there is a delay between the onset of symptoms and establishing a diagnosis, during which period the patients utilise healthcare resources. The linked routine data can provide information about the patients' visits to health care facilities during this symptomatic pre-diagnosis period, when the requirement for diagnostic investigations is often greatest. Within the SAIL data system, using the encrypted ALF, we can identify patients from a cohort of any particular disease who were diagnosed during a reference time; and link those with various datasets (e.g. GP data, inpatients hospital admissions, outpatients, A&E data etc.) to track their pre-diagnosis visits to healthcare facilities since the date the symptoms developed (as reported by the patients or established from the GP or A&E records). This allows one to compare the extent of health service utilisation, and therefore related costs, before and after the symptoms developed. In addition, one can also calculate the costs as a result of delayed diagnosis.

The linked healthcare analysis within SAIL need not be confined to deducing the extent of health service utilisation during the pre- and post-diagnosis illness periods but can additionally be performed to ascertain the size of the direct medical costs associated with the index illness that are incurred across different healthcare sectors. This is possible, for example, with the combined use of the cost figures included in the Trust Financial Return 2 (TFR2) accounts [9] that incorporate expenditures relating specifically to A&E attendances, inpatient admissions and outpatient contacts; and the cross-sector (i.e. primary care, secondary care, inpatients, outpatients, A&E etc) health services utilisation at the individual

level obtained from the linked data system. The SAIL system therefore not only allows the index event for a given condition to be identified but additionally introduces a longitudinal, temporal, dimension to the analysis as each of the healthcare sectors captured within SAIL can be searched for multiple years pre- and post-index healthcare event to determine the extent of health service utilisation and direct medical costs, made possible by the inclusion of an ALF within all of the SAIL datasets. This provides a more objective estimate of the actual costs of chronic conditions and any interventions.

4. Healthcare pathways and referral history

A retrospective analysis of the patient's healthcare history can identify the types of referrals to healthcare services made at different points in time, thus, giving an assessment of health service usage and recommendations for improving patient care pathways.

5. Profiling of patients

The linked SAIL data includes diverse sets of information, which enables profiling and stratification of patients relating to disease manifestation and severity, lifestyles, co-morbidities, and associated costs. Additionally, given that many chronic conditions have heterogeneous manifestations with a variable course and unpredictable episodes of exacerbation, the analysis can be carried out under several person stratification schemes based on severity of disease, various demographic attributes, and socio-economic conditions. This stratification will facilitate early targeting of interventions to patients at highest risk, thereby improving the cost-effectiveness ratio of these interventions.

An Example of a Patient with Ankylosing Spondylitis

As part of the Medical Research Council (MRC) Patient Research Cohort Initiative, a cohort of people with ankylosing spondylitis (AS), i.e. the Welsh population-based ankylosing spondylitis (PAS) cohort, has been developed using data collected from patient completed

questionnaires linked with routine data [5]. The study aims to recruit 1000 AS patients living in Wales and currently about 500+ AS patients are participating. We tracked the healthcare events of an arbitrarily chosen AS patient from the PAS cohort. The patient completed the questionnaire during the last week of September 2009, and was asked to recall the number of visits to the GP, outpatients (OP), inpatients (IP), A&E, and to various health professionals during the three months before the questionnaire completion date. The patient reported 4 GP visits, 1 OP visit, 1 IP visit, no A&E visit, and visited a rheumatologist, a radiologist, and a chiropractor once each. Distances to the healthcare facilities were 1.5 miles, 4 miles, and 3 miles for GP, OP, and IP, respectively. In each case the patient used their own car; and was accompanied by someone during the GP and IP visits. The patient also reported having taken pain reducing medicines (paracetamol, ibuprofen, and naproxen); having undergone an MRI scan and had blood and urine tests during the 3 months recall period.

Using the unique ALF, we tracked the patient's healthcare pathways through the routine data in SAIL system. Figure 2 plots the patient's actual healthcare events from the OP, GP, and A&E datasets for a 2 year period (i.e. August 2008 to August 2010), which represents the timeline approximately one year before and one year after the completion date of the questionnaire by the patient.

The linked routine data show 10 GP events, 2 OP visits and 1 A&E visit during the 3 month recall period. There is no IP visit recorded during this period. Therefore, the self-reported IP visit in the questionnaire may actually be an A&E visit, which would correlate with the routine data. Out of those 10 GP events, not all of them are physical visits by the patients, but may include any event (e.g. letter encounter, prescription collection, telephone conversation etc.). Further exploration of the GP read codes and descriptions for these 10 GP events yields information about medication, tests, and other GP related encounters, as shown in Table 1. Retrospective and prospective tracking of events reveals that there are 49 such GP events

during the 2 year period. There are no OP, IP, or A&E visits before or after the recall period, indicating the difficulty of extrapolating patient reported 3 month recall data in the questionnaire for a longer period (e.g. one year). However, going further back through the linked data system in terms of the timeline (not shown in the figure), it was found that the patient made 4 OP visits during April, June, and July of 2005; and 1 IP visit on the last week of March 1998.

The linked routine data therefore enables validation and clarification of the patient reported data; the retrospective and prospective tracking of the patient healthcare utilization and pathways; and the referral history. Such analysis makes it possible to deduce whether the anonymised individual in question was suffering any common co-morbidities in receipt of healthcare treatment prior to the occurrence of the reference event, whilst it also allows any frequent complications requiring medical attention in the days, months, years following the event to be identified. Subsequently, from the methodological perspective, this linkage system would add new dimensions and perspective into the traditional use of routine data in health related research, e.g. complement and enhance the results of RCTs, as a resource for clinical audits, and in various health impact assessment exercises. For example, the longitudinal routine data would allow an assessment of the impact of healthcare interventions on subsequent healthcare utilisation (e.g. A&E visits or hospital admission). An important limitation of solely relying on questionnaire data is that the healthcare events of interest may not occur within the limited recall period (e.g. 3 months), but just before the recall period or after the completion date of the questionnaire. This makes extrapolation of the questionnaire data for an extended period of time unreliable. The longitudinal linked routine data comes into aid in this respect.

The linkage of the questionnaire data from the PAS patients with the GP data as shown above enhances and helps make sense of the rich information obtained from the GP Read codes.

This constitutes a rich health history for these AS patients, for whom we will carry out patient pathway analysis from various clinical and economic aspects. In particular, in keeping with other AS cohorts, these patients had an average lag of about 8 years from symptom onset to AS diagnosis, on which we can conduct pre- and post diagnosis analyses.

Again, a matrix of traits based on the PAS questionnaire information linked with the SAIL data system will help profiling of AS patients for health and other related interventions. Table 2 summarises the types of information gathered through the PAS questionnaires, which subsequently could be linked with the routine data sources as well as various demographic, socio-economic, and environmental attributes of the patients.

The use of HERALD methodology can stratify groups of patients to identify the early characteristics of patients who subsequently develop severe disease, thus, enabling these patients to be targeted with early aggressive therapy in order to prevent severe damage and need for surgery. This profiling can be used to estimate the potential resource savings of treating those patients with patterns of disease suggestive of the development of a severe outcome. All these will directly affect patient care for AS in terms of informing NHS service provision and NICE guidelines for the use of expensive biological therapy, and informing the process of assessment of cost-effectiveness. Intuitively, the methods developed for the PAS cohort can be extrapolated to be used in other chronic disease conditions; thus improving patient care for all those conditions.

Discussion

Linked routine data in the SAIL and other similar databanks provides many opportunities for enhanced healthcare research and allows evaluation of impacts beyond the limited primary outcomes of interventional studies. The growing SAIL databank already holds over a billion anonymised records from various databases, which can be anonymously linked at the

individual record level. The combination of routine data with information from patients and RCTs allows the validation of real-life data and its application for clinical research. These linkable databases provide factual and continuous information with rich clinical and non-clinical details, which offers wide ranging opportunities in the realm of conducting evaluative research, clinical epidemiology, trial recruitment, genetic research, basic research of biological markers, stratified medicine, post-trial surveillance, risk assessment, service delivery evaluation, resource use, decision analysis, identification of early disease predictors, and the identification of subjects for prospective studies [5, 10, 11]. This model also offers the opportunity for post-marketing surveillance and pharmacovigilance of new expensive, and often potentially dangerous, healthcare interventions in real-life settings. Complementing this resource with targeted health economic analysis, as proposed in the HERALD methodology, offers a unique opportunity to deliver the level of health economic data required to evaluate and drive forward modern healthcare services.

Conclusion

The linkage of routine data, patient completed questionnaires and trial data offers unique opportunities for enhanced health economic analysis, including assessment of the validity, reliability and generalisability of health economic data not possible through the use of traditional isolated datasets.

.

Authors' contributions

All authors were involved in the design of the HERALD protocol. The first draft of the paper was written by MJH, SB, and SS. Subsequent drafts were amended and finally approved by all the authors.

Acknowledgements/Funding

Funding for the PAS cohort was provided by Medical Research Council (MRC) and National Institute for Social Care and Health Research (NISCHR).

References

[1] McLean G: **A review of the use of routine data on CHD in general practice research in Scotland.** *Scottish Medical Journal*. 2005. **50**(3) pp.109-113.

[2] Hansell A, Aylin P: Routine data and health impact assessment – a review of epidemiological studies of socio-economic influence on health and evaluation of outcome indicators derived from routine health data for health impact assessment. *Report to the Department of Health*. The Small Area Health Statistics Unit, Dept of Epidemiology and Public Health, Imperial College School of Medicine at St Mary's, March 2000.

[3] Lyons RA, Jones KH, John G, Brooks CJ, Verplancke JP, Ford DV, Brown G, Leake K: **The SAIL databank: linking multiple health and social care datasets.** *BMC Medical Informatics and Decision Making* 2009, **9**:3.

[4] Ford DV, Jones KH, Verplancke JP, Lyons RA, John G, Brooks CJ, Thompson S, Bodger O, Couch T, Leake K: **The SAIL databank: building a national architecture for e-health research and evaluation.** *BMC Health Services Research* 2009, **9:**157.

[5] Atkinson MD, Brophy S, Siebert S, Gravenor MB, Phillips C, Ford DV, Jones KH, Lyons RA: **Protocol for a population-based Ankylosing Spondylitis (PAS) cohort in Wales**. *BMC Musculoskeletal Disorders* 2010, **11**:197

[6] Graves N, Walker D, Raine R, Hutchings A, Roberts JA: **Cost data for individual patients included in clinical studies: no amount of statistical analysis can compensate for inadequate costing methods.** *Health Economics*, John Wiley & Sons, Ltd., vol. **11**(8), pages 735-739, 2002.

[7] Halliday RG, Darba J: **Cost data assessment in multinational economic evaluations: some theory and review of published studies.** *Appl Health Econ Health Policy.* 2003; **2**(3):149-55.

[8] Drummond MF , Sculpher MJ, Torrance GW, O'Brien BJ, Stoddart GL: *Methods for the Economic Evaluation of Health Care Programmes***.** OUP Oxford; 3 edition (2 Jun 2005).

[9] Department of Health: **NHS Trust financial returns form 2 (TFR2) specialty-specific costs per in patient day for hospital inpatient care and day case care**. London: HMSO, 2000.

[10] Black N: **High-quality clinical databases: breaking down barriers.** *Lancet* 1999, **353**:1205-1206.

[11] Tierney, W. M. and McDonald, C. J.: **Practice databases and their uses in clinical research**. *Statistics in Medicine*, **10**: 541–557. doi: 10.1002/sim.4780100406, 1999.

[12] Garrett S, Jenkinson T, Kennedy LG, Whitelock H, Gaisford P, Calin A: **A new approach to defining disease status in ankylosing spondylitis: the Bath Ankylosing Spondylitis Disease Activity Index.** *J Rheumatol* 1994, **21**:2286-2291.

[13] Calin A, Jones SD, Garrett SL, Kennedy LG: **Bath ankylosing-spondylitis functional index**. *British J Rheumatol* 1995, **34**:793-794.

[14] Cheung K, Oemar M, Oppe M, Rabin R: **EQ-5D user guide. Basic information on how to use EQ-5D Rotterdam**: EuroQol Group 2009.

[15] Lerner D, Amick BC III, Rogers WH, Malspeis S, Bungay K, Cynn D: **The Work Limitations Questionnaire: a self-administered instrument for assessing on-the-job work disability**. *Med Care 2001*, **39**:72-85.

[16] WPAI general information, Reilly Associates. [http://www.reillyassociates.net/].

[17] Reilly MC, Gooch KL, Wong RL, Kupper H, van der Heijde D: **Validity, reliability and responsiveness of the Work Productivity and Activity Impairment Questionnaire in ankylosing spondylitis.** *Rheumatology* 2010, **49**:812-819.

[18] The International Physical Activity Questionnaire. [http://www.ipaq.ki.se/ipaq.htm].

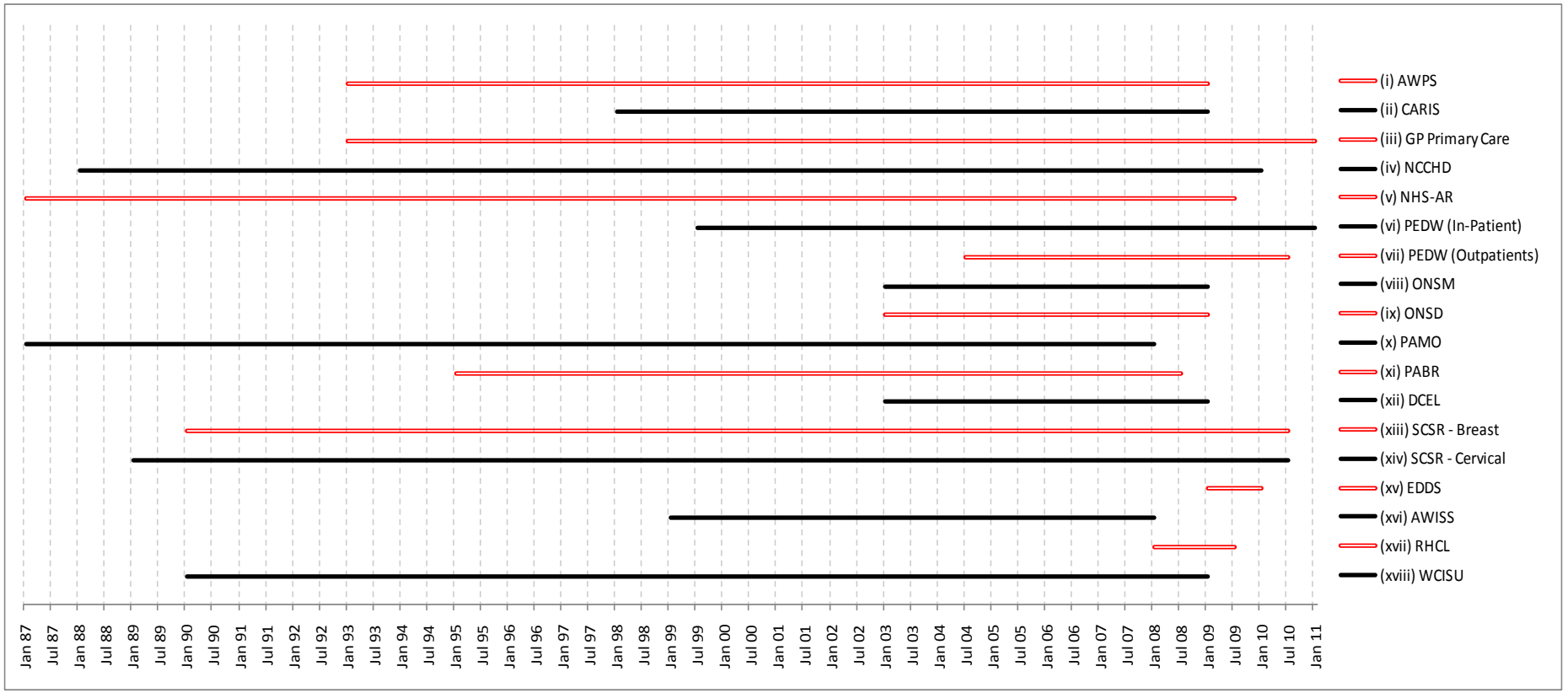
[19] Markland T, Tobin V: **A Modification to the Behavioural Regulation in Exercise Questionnaire to Include an Assessment of Amotivation**. *J Sport Exercise Psychol* 2004, 26:191-196.

[20] Buysse DJ, Reynolds CF III, Monk TH, Berman SR, Kupfer DJ: **The Pittsburgh Sleep Quality Index: a new instrument for psychiatric practice and research.** *Psychiatry Research* 1989, **28**:193-213.

[21] Zigmond AS, Snaith RP: **The hospital anxiety and depression scale**. *Acta psychiatr Scand* 1983, **67:**361-370.

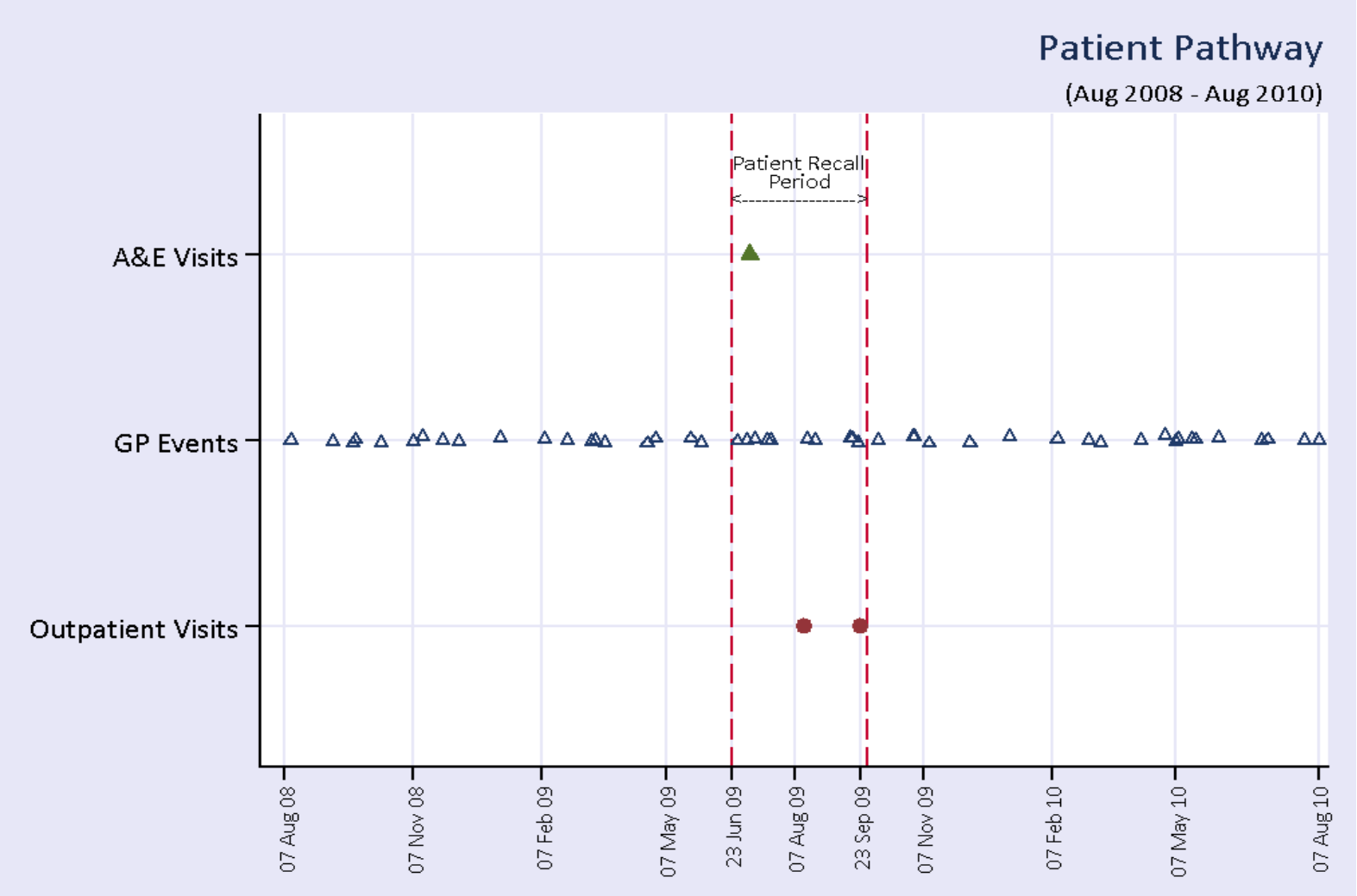
[22] Snaith RP: **The hospital anxiety and depression scale**. *Health and Quality of Life Outcomes* 2003, **1**:29.

Figure 1: Timelines for the Availability of Datasets in the SAIL System



Legend: (i) **AWPS:** All Wales Perinatal Survey; (ii) **CARIS**: Congenital Anomaly Register and Information Service for Wales; (iii) **GP Primary Care:** GP Primary Care Audit+; (iv) **NCCHD**: Child Health data; (v) **NHS-AR**: NHS Administrative Register; (vi) **PEDW (In-Patient)**: NHS Hospital Admission; (vii) **PEDW (Outpatients)**: NHS Hospital Outpatients; (viii) **ONSM**: Office for National Statistics- Deaths; (ix) **ONSD**: Office for National Statistics- Births; (x) **PAMO**: Pathology – ABM Trust; (xi) **PABR**: Pathology – Abertawe Bro Morgannwg; (xii) **DCEL**; Pre16 Education Attainment; (xiii) **SCSR-Breast**: Screening Service - Breast; (xiv) **SCSR-Cervical**: Screening Service - Cervical; (xv) **EDDS**: Emergency Department Dataset; (xvi) **AWISS**: All Wales Injury Surveillance System; (xvii) **RHCL**: Rheumatology Register; (xviii) **WCISU:** Welsh Cancer Intelligence Surveillance Unit – Cancer Registry Data. The figure shows the datasets that are on SAIL (or in the process of loading on to SAIL) as at November 02, 2010.

Figure 2: Healthcare Pathway of an AS Patient



Legend: Each shape indicates a separate event (date) captured in the relevant datasets. The X-axis captures events for a period of two years (i.e. 730 days). Therefore, the GP event markers appear almost overlapped when events occurred in two consecutive dates (e.g. 14th and 15th September). The questionnaire completion date is 26th September, 2009. The three month patient recall period is captured by the two vertical dashed lines, indicating the dates 23rd June and 26th September, 2009. There are 10 GP events during this time interval, including the overlapped one.

Table 1: Patient's GP Read codes during the 3 month recall period

GP Event Date	(sl) GP Read Codes: Descriptions	Comment
23 June	(i) fh1k: PREMIQUE 0.625mg/5mg tablets; (ii) b211: BEMDROFLUAZIDE 2.5 mg tablets; (iii) b211: BENDROFLUMETHIAZIDE 2.5 mg tablets; (iv) bxd2: SIMVASTATIN 20mg tablets; (v) Bd35: ATENONOL 50mg tablets	GP visit /Prescription collection
06 July	1969: Abdominal pain	System reporting of A&E admission on 06 July
07 July	Discharge Summary	System reporting of A&E discharge summary (ref: 06 July A&E admission)
20 July	(i) fh1k: PREMIQUE 0.625mg/5mg tablets; (ii) b211: BEMDROFLUAZIDE 2.5 mg tablets; (iii) b211: BENDROFLUMETHIAZIDE 2.5 mg tablets; (iv) bxd2: SIMVASTATIN 20mg tablets; (v) Bd35: ATENONOL 50mg tablets; (vi) j2ck: NAPROXEN 250 mg e/c tablets; (vii) 246..: O/E blood pressure; (viii) di21: PARACETAMOL 500mg; (ix) 8B4: Repeat Prescription.	GP Visit
24 July	(i) 52…: Plain Radiography; (ii) 52….:Plain X-Ray	GP visit
17 August	(i) bd35: ATENONOL 50mg tablets; (ii) b211: BEMDROFLUAZIDE 2.5 mg tablets; (iii) bxd2: SIMVASTATIN 20mg tablets; (iv) 8CB..: Had a chinwag with patient; (v) 8CB..: Had a discussion with patient.	GP visit
18 August	9N4..: Failed encounter	No visit
14 September	(i) b211: BEMDROFLUAZIDE 2.5 mg tablets; (ii) b211: BENDROFLUMETHIAZIDE 2.5 mg tablets; (iii) bxd2: SIMVASTATIN 20mg tablets; (iv) Bd35: ATENONOL 50mg tablets (v) j2ck: NAPROXEN 250 mg e/c tablets	GP visit/prescription collection
15 September	(i) j2ck: NAPROXEN 25 mg e/c tablets	Prescription collection
23 September	(i) 9N36: Letter from specialist; (ii) Letter from consultant	No visit

Table 2: PAS Questionnaire Contents

Questionnaire (Time Interval)	Summary of Information
Baseline (0)	Co-morbidities, family history, age of diagnosis and first symptoms, disease activity [12], function [13], Quality of life (EQ-5D) [14], and visits to health professionals
Not at work (3 Months)	Previous occupation, activity impairment questionnaire
At work (3 months)	Work questionnaire, including information about current and previous occupation, activity impairment and work limitations questionnaire (WLQ) [15], work productivity and activity impairment questionnaire (WPAI-SHP) [16, 17]
AS costs (9 months)	AS Cost questionnaire including detailed patient-level information about visits to health care facilities, professionals, AS related pathology and other tests, other conditions, medications, costs of various aspects of treatment and disability
Exercise and Fatigue (15 months)	International Physical Activity Questionnaire (IPAQ) [18], disease activity, function, Behavioural Regulation in Exercise Questionnaire (BREQ-2) [19], Pittsburgh Sleep Quality Index [20], and the Hospital Anxiety and Depression Scale [21, 22])
Medication (0,3,6,9,12,15 Months)	Medication

Legend: AS patients in the PAS cohort were asked to consent to completing questionnaires either online if they have internet access, or by post. A website has been developed to give access to the questionnaires (http://www.ashealth.co.uk/).