Validation of the Hospital Episode Statistics Outpatient Dataset in England

Joanna C. Thorn¹ · Emma Turner¹ · Luke Hounscome² · Eleanor Walsh¹ ·
Jenny L. Donovan¹ · Julia Verne² · David E. Neal³ · Freddie C. Hamdy⁴ ·
Richard M. Martin¹ · Sian M. Noble¹

Published online: 19 September 2015
© Springer International Publishing Switzerland 2015

Abstract

Objectives The Hospital Episode Statistics (HES) dataset is a source of administrative ‘big data’ with potential for costing purposes in economic evaluations alongside clinical trials. This study assesses the validity of coverage in the HES outpatient dataset.

Methods Men who died of, or with, prostate cancer were selected from a prostate-cancer screening trial (CAP, Cluster randomised trial of PSA testing for Prostate cancer). Details of visits that took place after 1/4/2003 to hospital outpatient departments for conditions related to prostate cancer were extracted from medical records (MR); these appointments were sought in the HES outpatient dataset based on date. The matching procedure was repeated for periods before and after 1/4/2008, when the HES outpatient dataset was accredited as a national statistic.

Results 4922 outpatient appointments were extracted from MR for 370 men. 4088 appointments recorded in MR were identified in the HES outpatient dataset (83.1 %; 95 % confidence interval [CI] 82.0–84.1). For appointments

occurring prior to 1/4/2008, 2195/2755 (79.7 %; 95 % CI 78.2–81.2) matches were observed, while 1893/2167 (87.4 %; 95 % CI 86.0–88.9) appointments occurring after 1/4/2008 were identified (p for difference <0.001). 215/370 men (58.1 %) had at least one appointment in the MR review that was unmatched in HES, 155 men (41.9 %) had all their appointments identified, and 20 men (5.4 %) had no appointments identified in HES.

Conclusions The HES outpatient dataset appears reasonably valid for research, particularly following accreditation. The dataset may be a suitable alternative to collecting MR data from hospital notes within a trial, although caution should be exercised with data collected prior to accreditation.

Key Points for Decision Makers

The English Hospital Episode Statistics (HES) outpatient dataset appears reasonably valid for research purposes.

Researchers wishing to use the HES outpatient dataset should bear in mind its limitations, including lower coverage pre-accreditation, poor record of diagnostic information throughout the dataset and variation in the data collected over time.

To use the HES outpatient dataset in a trial context, it will be necessary for the scope of the economic analysis to include all resource use, as HES does not include sufficient data to accurately identify the reason for an appointment.

We have provided a straightforward methodology by which the validity of routinely collected outpatient datasets in any context (overseas or for different patient groups) can be tested.

✉ Joanna C. Thorn
joanna.thorn@bristol.ac.uk

¹ School of Social and Community Medicine, University of Bristol, Canynge Hall, 39 Whatley Road, Bristol BS8 2PS, UK

² Public Health England Knowledge and Intelligence Team (South West), 2 Rivergate, Temple Quay, Bristol BS1 6EH, UK

³ Department of Oncology, Addenbrooke's Hospital, University of Cambridge, Box 279 (S4), Hills Road, Cambridge CB2 0QQ, UK

⁴ Nuffield Department of Surgical Sciences, University of Oxford, Old Road Campus Research Building, Oxford OX3 7DQ, UK

1 Introduction

Routinely collected ‘big data’ sets (defined as large complex datasets requiring analysis with specialist hardware and software rather than commonly available tools [1]) in healthcare have the potential to contribute substantially to research efforts, both to inform standalone studies and to contribute data to randomised controlled trials (RCTs) [2, 3]. Barriers to the effective implementation of big datasets in research, however, include concerns over the accuracy of the data (in terms of both content and coverage), the time lag between an event and its availability in the dataset, and concerns over the technical capability of the provider organisation to manage large and complex datasets [4, 5]. Despite these issues, the use of health administrative data for costing purposes has the potential to improve research efficiency in health economics, and health economists are being encouraged to use them. As such data are gathered for the purpose of managing health systems, researcher requirements are not necessarily met. Information on whether routine data sources are accurate, and how they compare with other methods of obtaining resource-use data, is required to interpret results meaningfully.

The Hospital Episode Statistics (HES) dataset is a routine ‘big data’ source for which research use may be feasible [6]. HES is a national data warehouse that contains details of all secondary care contacts in England; datasets covering inpatient admissions, outpatient appointments, and accident and emergency visits are published annually. The HES outpatient dataset, on which we focus in this work, includes records of all National Health Service (NHS)-related outpatient activity in England, including private patients who have been treated in an NHS hospital, patients living outside (but treated within) England and patients whose treatment has been funded by the NHS but carried out in independent institutions. Data recorded include dates of appointments, type of professional seen and geographical information; a full list of the fields can be found in the HES outpatient data dictionary [7]. Outpatient data have been recorded from 2003 onwards; the data were first released on an ‘experimental’ basis in 2006, but have been accredited as a national statistic since 2008. The number of recorded appointments is expanding year on year, and reached over 100 million records for the year 2013–2014 [8].

The HES dataset is recorded for the purposes of reimbursement within the health system, and its validity for research purposes has not been thoroughly tested. HES analysts compile the dataset from submissions made by more than 200 trusts and over 500 private organisations, resulting in potential for variation in accuracy. Errors can arise in healthcare records because the data are largely

entered by hand, and simple human error (such as typographical errors) occurs. In addition, if data are not entered by the clinicians themselves, there is a requirement for the data entry clerk to interpret data that might be presented in numerous different ways (for example, applying handwritten operational codes). While validity of the submissions has improved over time, the Health and Social Care Information Centre acknowledge that there is still room for improvement, and data quality issues are actively monitored [8]. HES undergoes a fully automated data cleaning system according to published rules [9], and therefore has a good level of internal validity. However, the external validity of HES has been called into question since its inception, with concerns raised over the lack of clinician input in the early days [10].

Various aspects of the external validity of HES data have been studied, although mostly in the context of the inpatient dataset. Validation studies have shown that HES accuracy can outperform voluntary specialist databases in identifying patient volume [11, 12]. Poor levels of accuracy were found for primary diagnosis codes for inpatient stays in earlier years [13], although accuracy has improved more recently and is expected to continue to improve [14]. It is likely that inpatient data are recorded more accurately than outpatient data; inpatient care is a bigger cost driver for reimbursement, with the concomitant increase in incentive to code an event accurately, and more fields are typically required to generate the HRG required for reimbursement. For outpatient events, administrators are dealing with a large number of appointments and patients, increasing the scope for error and for missing data. Inpatient data have also been collected, and the procedures refined, for a considerably longer time than outpatient data [5].

Both the accuracy of the stored information and the completeness of the records are of concern. Early estimates of the validity of HES outpatient coverage compared volumes recorded in HES with volumes expected through other routine reports, and suggested that coverage of outpatient appointments was close to 100 % at an aggregate national level; however, it was acknowledged that there was some variation between under and over reporting in HES by area, trust and specialty [15]. A review of papers assessing the validity of inpatient data from HES for England and from the Patient Episode Database for Wales (PEDW) concluded that it was unlikely that the databases are able to capture all events of relevance [10]. Research aiming to validate self-reported strokes in a cohort study found that a number of these strokes were recorded in the patient’s medical records (MR), but did not appear in HES, indicating that HES coverage may not be complete [16].

HES has considerable potential for identifying resource use in economic evaluation and costing studies without the

need to access individual notes or burden the patient with requests for data. Examples of the use of the inpatient dataset for health economic purposes to date include investigating the by-hospital variation in costs for different interventions [17] and conducting a cost-benefit analysis for an immunisation programme [18]. However, although the outpatient dataset has been used in research (e.g. [19–21]), and for costing purposes in particular [22], we are not aware of any health economic RCT-based applications to date.

In this study, we assessed the completeness of the HES outpatient dataset in the context of male patients diagnosed with prostate cancer by comparing recorded appointments with data extracted from MR. From this, we aimed to establish whether the HES outpatient dataset is suitable for use for health economic research purposes.

2 Methods

2.1 Sample

The CAP trial (Cluster randomised trial of PSA testing for Prostate cancer, ISRCTN92187251) is assessing the effectiveness and cost effectiveness of prostate-specific antigen (PSA) testing [23]. General practices are randomised to participate in the intervention arm (the ProtecT trial, ISRCTN20141297) or the control arm. In the ProtecT arm, men are invited to be tested for the presence of prostate cancer through population-based PSA testing [24]. In the control arm, men receive standard NHS care (NHS prostate cancer risk management programme). Individuals in the CAP trial are men aged 50–69 years who are registered with one of the 573 CAP study practices.

This sub-study uses resource-use data from a sample of all men in CAP associated with one of the seven centres based in England (Sheffield, Birmingham, Bristol, Cambridge, Leicester, Newcastle and Leeds). The men met the following inclusion criteria: had the presence of prostate cancer confirmed; died either of prostate cancer or of other causes; had at least one outpatient appointment recorded in their MR between April 2003 and January 2012; and their MR had undergone a complete review. This latter requirement renders the sample ‘quasi-random’ because men were arbitrarily selected for review from lists supplied by the cancer registry; groups of men based in any particular hospital will tend to be reviewed together. Men entered the study on the date that their practice joined CAP or April 2003 (whichever was later), and left at death or January 2012 (whichever was earlier). Descriptive statistics of the men were derived, including age and index of multiple deprivation based on the general practitioner practice postcode (lower layer super output area, an area

level measure of deprivation [25]). The cause of death was established by a committee through a stringent assessment process for a subset of the men [26].

2.2 Outpatient Resource Use

A detailed review of MR is undertaken for men in CAP who have received a diagnosis of prostate cancer during their lifetime, or have died with prostate cancer or bone metastases cited on the death certificate. Ethical approval for review of MR for men with prostate cancer (provisional on the men giving their consent) was provided by the Trent Medical Research and Ethics Committee [05/MRE04/78]. Provided no objection to the use of his MR for research purposes was recorded during the man’s lifetime, the Patient Information Advisory Group granted support for reviewing MR of men who potentially died of a cause related to prostate cancer before we could gain their consent [PIAG 1-05(f)/2006]. The MR reviews were undertaken by a team of eight researchers who had no knowledge of the data captured in HES, using a standardised pro forma. These researchers are experienced in extracting medical data and undergo regular training events to ensure a consistent approach. Details of all outpatient appointments relating to prostate cancer were extracted from hospital-based records, including appointment date, reason for appointment, department visited and type of practitioner. Summary characteristics of the appointments were derived.

HES outpatient data for all appointments (i.e. not just those relating to prostate cancer) between April 2003 and January 2012 were obtained for the same men. Patients were matched in the HES outpatient dataset using NHS number and sex, with date of birth, forename, surname and postcode used to supplement the process where necessary. Details of appointment date, information on attendance, diagnoses, operations, specialties and type of practitioner were available. The dataset also included the healthcare resource group (HRG) assigned to each appointment, spanning versions 3.5 and 4.0: HRGs are defined as groups of clinically similar events identified as consuming similar levels of resources [27] and are, therefore, of interest for health economic purposes. HRG version 3.5 includes about 600 codes each comprising three characters; five-character HRG version 4.0 codes, introduced in April 2007, number around 1400 allowing a far more granular approach, although outpatient appointments are frequently coded to the subchapter labelled ‘WF’, which contains eight codes only. Characteristics of the HES outpatient dataset were examined. To assess whether specialty codes (e.g. 101 for urology) could be used as a proxy for identifying condition-specific appointments, appointments carried out in urology, oncology or palliative departments (i.e. specialties

considered likely to be associated with prostate cancer treatment) were examined.

2.3 Analysis

Statistical analysis was carried out using Stata 12 [28]. Appointments identified through the MR review process were treated as a pseudo-gold standard, and matching appointments (based on appointment date) were sought in the HES outpatient dataset using Stata's merging functionality. Proportions matching were found for periods before and after 1/4/2008, when the HES outpatient dataset was accredited as a national statistic. A match to an appointment that the patient did not attend (DNA) was accepted as a valid match, as it showed evidence that the event had been captured in HES. Chi squared tests were performed to test whether the proportions differed significantly; *p* values and 95 % confidence intervals (CIs) were derived. The rate of matching was also found for appointments of men who died pre- and post-accreditation. Sensitivity analyses were performed by allowing a tolerance of ± 2 days in the appointment date, by excluding appointments associated with men for whom there was no record in HES and by excluding appointments that could potentially have taken place in hospices. The rate of matching over the whole study period was derived for each region, and for both pre- and post-accreditation periods; 95 % CIs were calculated, and a Chi squared test assessed whether the best and poorest performing regions differed significantly.

3 Results

Results are reported with reference to relevant aspects of guidelines for reporting validation studies of health administrative data [29].

3.1 Sample

370 men dying between April 2003 and October 2012 met the inclusion criteria; a summary of their characteristics is given in Table 1. 1590 person years of data were included in the analysis. The men in this sub-study came from slightly more deprived areas than those in CAP as a whole [23].

3.2 Medical Records Review Dataset

4922 outpatient appointments occurring between April 2003 and January 2012 were extracted from MR for the 370 men. A summary of the characteristics of the appointments is given in Table 2. The majority of

Table 1 Characteristics of the included men

No. of men included in study	370
Mean (SD) length of time in study (years)	4.3 (1.8)
Mean (SD) age at entry to study	63.6 (5.1)
Mean (SD) age at diagnosis	65.2 (5.3)
Mean (SD) age at death	67.5 (5.4)
Mean (SD) time between diagnosis and death (days)	816 (573)
Number (percentage) of men definitely or probably dying of prostate cancer according to the cause of death committee (<i>n</i> = 324)	207 (63.9)
Mean (SD) English IMD score	25.9 (17.4)
No. (%) of men in each centre	
Sheffield	106 (28.7)
Newcastle	82 (22.2)
Bristol	57 (15.4)
Leicester	40 (10.8)
Leeds	38 (10.3)
Cambridge	27 (7.3)
Birmingham	20 (5.4)

All men died of prostate cancer or other causes

IMD index of multiple deprivation, SD standard deviation

appointments (4713, 95.75 %) were held in urology, oncology or palliative departments.

3.3 Hospital Episode Statistics Dataset

The HES outpatient dataset included 12,154 appointments between April 2003 and January 2012; the large difference compared with MR review (4922) arises because HES data cover appointments for all reasons, not just those related to prostate cancer. All records have the appointment date completed. However, only 110 (0.9 %) have detailed codes to describe the diagnosis, and only 820 appointments (6.7 %) have operation codes that are not defined as errors or missing. By contrast, specialty is completed well with only 52 appointments (0.4 %) having both the main specialty (i.e. the specialty under which the health professional was contracted to work) and the treatment specialty (i.e. the specialty under which the health professional was actually working) missing; 7452 appointments (61.3 %) were identified as being carried out under urology, oncology or palliative treatment specialties. HRG version 4 codes are consistently present after 1/4/2009, with 91.2 % of appointments coded to the WF subchapter (2782/3051), 6.0 % (184/3051) to other HRGs and 2.8 % (85/3051) to error codes. First attendance codes (that define whether the appointment was a follow-up or not) were well completed, with only one appointment (0.01 %) missing this information; 1895 appointments (15.6 %) were defined as being first face-to-face consultations.

Table 2 Characteristics of outpatient appointments according to medical records review

Characteristic	No. (%) of appointments
Centre	
Sheffield	1753 (35.6)
Newcastle	864 (17.6)
Bristol	699 (14.2)
Leicester	441 (9.0)
Leeds	512 (10.4)
Cambridge	395 (8.0)
Birmingham	258 (5.2)
Appointment led by	
Nurse	1129 (22.9)
Consultant	3721 (75.6)
Unknown	72 (1.5)
Department of appointment	
Urology	2512 (51.0)
Surgery	14 (0.3)
Oncology	1829 (37.2)
Joint urology/oncology	282 (5.7)
Medicine	11 (0.2)
Palliative	90 (1.8)
Care of the elderly	6 (0.1)
Other	178 (3.6)
Reason for appointment	
Routine follow-up	2758 (56.0)
Receive hormone treatment	593 (12.0)
Discuss treatment	402 (8.2)
First consultation	316 (6.4)
Complications	172 (3.5)
Palliative care	120 (2.4)
Other	561 (11.4)

3.4 Event Identification

4088/4922 appointments recorded in MR were identified in the HES outpatient dataset (83.1 %; 95 % CI 82.0–84.1) based on an exact date match. Seven of the appointments in

MR review (0.1 %) matched to appointments in HES for which it is recorded that the patient ‘did not attend’ (DNA). Allowing a ± 2 day tolerance for the appointment date resulted in a slight improvement to 4173/4922 matches (84.8 %; 95 % CI 83.8–85.8; $p = 0.02$). For appointments occurring when the dataset was considered experimental (prior to 1/4/2008), 2195/2755 (79.7 %; 95 % CI 78.2–81.2) matches were observed, while 1893/2167 (87.4 %; 95 % CI 86.0–88.9) appointments occurring after 1/4/2008 were identified ($p < 0.001$). 97/370 men (26.2 %) died prior to 1/4/2008; 760/970 of their appointments matched (78.4 %), compared with 3328/3952 (84.2 %) for men dying after 1/4/2008 ($p < 0.001$). Regional variation in the rate of matching was observed with the best performing centre (Cambridge) achieving a match rate of 351/395 (88.9 %; 95 % CI 85.8–92.0) and the least well performing centre (Leeds) achieving a match rate of 391/512 (76.4 %; 95 % CI 72.7–80.1, p for difference Cambridge vs. Leeds < 0.001). The rate of matching for all regions, stratified by pre- and post-accreditation periods, is presented in Table 3.

215/370 men (58.1 %) had at least one appointment in MR review that was unmatched in HES, 155 men (41.9 %) had all their appointments identified and 20 men (5.4 %) had no appointments identified in HES (two of these men had appointments in HES, but these were not matched with any of the MR review appointments). If the 18 men (4.9 %) for whom there is no record at all in HES are excluded from the analysis, the proportion of matching appointments increases to 4088/4674 (85.7 %; 95 % CI 86.5–88.4). Thirty-three appointments were identified as potentially having occurred in hospices; excluding them results in 4084/4889 appointments matching (83.5 %; 95 % CI 82.5–84.6).

4 Discussion

This study is the first to look at the completeness of the HES outpatient dataset by comparing it with another source of resource-use data. Most appointments extracted during

Table 3 Regional variation in the rate of matching appointments

Centre	Percentage matching overall (95 % CI)	Percentage matching post-2008 (95 % CI)	Percentage matching pre-2008 (95 % CI)
Sheffield	87.8 (86.3–89.3)	87.7 (85.5–89.9)	87.9 (85.8–90.0)
Newcastle	78.6 (75.9–81.3)	82.9 (79.0–86.9)	75.7 (72.0–79.4)
Bristol	79.7 (76.7–82.7)	90.5 (86.6–94.4)	74.7 (70.8–78.6)
Leicester	81.2 (77.5–84.8)	88.9 (84.6–93.2)	74.4 (68.7–80.0)
Leeds	76.4 (72.7–80.1)	83.5 (79.1–87.8)	67.9 (62.0–73.9)
Cambridge	88.9 (85.8–92.0)	94.0 (90.5–97.5)	84.4 (79.4–89.3)
Birmingham	82.6 (77.9–87.2)	88.1 (81.8–94.5)	79.0 (72.6–73.9)

CI confidence interval

the MR review procedure were also identified in the HES outpatient dataset based on an exact date match. Allowing a tolerance of ± 2 days in the date resulted in slightly better matching. The proportion of appointments identified rose significantly following accreditation of the HES outpatient dataset in April 2008.

The HES dataset had a lower proportion of appointments carried out in prostate-cancer related specialties (61.3 % compared with 95.8 %) because MR review explicitly sought only prostate-cancer related appointments, the majority of which occur in a small range of specialties. The HES outpatient dataset also had a higher proportion of appointments identified as being first appointments (15.6 % compared with 6.4 %); this could have arisen because the men attended multiple first appointments for conditions other than prostate cancer.

There is substantial confusion over the difference between day cases and outpatient appointments within the NHS [30]. It is possible that the appointments for which no match was found in the HES outpatient dataset could appear as day cases in the HES inpatient dataset; for example, patients undergoing catheterisation could have been recorded as either outpatient or day case events. The 18 men for whom no appointments at all appeared in HES could have failed to be matched on their NHS numbers if either HES or the CAP trial held incorrect data; however, the increase in matching if these men were excluded from the analysis was relatively small and did not suggest strong evidence of a difference. Regional variation in matching could have arisen from different methods of producing the data. Despite the attempted standardisation of HES entries, comparison of HES between NHS trusts may be problematic, as the methods by which the data are produced vary according to the size of the trust and the amount of contact between clinicians and coders [31].

In common with other studies, we found that coverage of events in HES was not complete [10, 16] when compared with an alternative data source. However, in comparison with studies of inpatient data, we found a relatively small amount of missing data. It is likely that this loss will be offset by the fact that HES will also contain appointments that would be missed by MR review (if, for example, the records are held in a hospital out of the area covered by the researchers) and that HES contains information about whether the patient attended an appointment or not. This suggests that the HES outpatient dataset is a suitable alternative to reviewing MR to identify a patient's resource use. Although this study was carried out in the context of an RCT, the finding that the HES outpatient dataset is valid for research is not necessarily restricted to RCTs and has relevance to any study in which a patient's resource use is required. Data quality in the HES outpatient dataset is

actively monitored and reported [8], and is likely to improve further over time.

In this study, MR review was treated as the gold standard against which the HES data were tested. The data extraction process was conducted rigorously, with researchers undergoing regular training events. Because only details of appointments that the researchers identified as being related to prostate cancer were extracted, whilst HES contains details of all appointments but does not reliably identify the reason, it was not possible to co-test the validity of the MR review dataset. The important consideration from a health economic point of view is that events are neither under nor over estimated; we cannot state for certain whether HES or MR review represents the gold standard as both have weaknesses.

An open question for all economic evaluations is whether resource-use data should be restricted to disease-specific events, or more broadly sought for all healthcare events [32]. Including all resource use increases the research burden as greater quantities of data must be collected, and can lead to overinflated estimates if particularly costly events occur that are unrelated to the study condition. However, it is not always straightforward to accurately identify events associated with the study disease; for example, if a trial is studying an intervention for diabetes, it may not be possible to identify whether a visit to hospital for a fall is related to the diabetes or not. For this reason, trials including all resource use potentially have less scope for bias. Decisions on all-cause or disease-specific resource use are taken on a trial-by-trial basis; however, if health economists plan to use the HES outpatient dataset in a trial, it will be necessary to include all resource use, as HES does not include sufficient data to accurately identify the reason for an appointment. The use of specialty as a proxy is unlikely to be effective; in the MR review, we identified a number of appointments believed to be related to prostate cancer that were carried out under different specialties. Equally, basing an analysis on specialties considered likely to be related to prostate cancer (urology, oncology, palliative care) will include care related to other conditions (e.g. bladder cancer). Hospital consultants identified the lack of diagnosis record in the outpatient dataset as a priority for future improvements in HES; over 90 % agreed that they would code diagnoses if appropriate tools were available [33], therefore this limitation may change in the future.

Advantages of using HES data in an economic evaluation in preference to manually extracting data from MR include the fact that HES covers all hospitals in England rather than the more limited number that can be accessed in person. This allows the capture of events (such as unexpected complications) taking place outside a patient's home territory, and potentially reduces the amount of

missing data. The single point of access avoids issues with hospitals having varying access requirements, and avoids the requirement to obtain multiple research and development approvals. As HES data are routinely returned, there is no additional burden on the hospital associated with supervising access for manual data extraction, and retrieval of records is substantially less time consuming. However, researchers wishing to use the HES outpatient dataset should bear in mind its limitations. The dataset is region specific, incorporating only appointments that occur in England; appointments for English patients occurring elsewhere are not included. In particular, appointments occurring in Wales are not available through HES: an integrated system including both Welsh PEDW data and English HES data would be valuable for researchers. While the requirement to use codes rather than free text to describe an event produces datasets that are more readily analysable, it also leads to a loss of information. Clinical coders are skilled employees; however, free text medical notes can be hard to interpret even by experts [33]. Care should be taken when accessing HES data over multiple years because methods of data collection, variables and reporting standards alter over time. Researchers should also note the limitations with regard to missing data and known problems published on the HES website [34]. As is common with routine datasets, there is a time lag between the event and the availability of the data describing that event in HES; there is currently approximately a 6-month delay [8].

Public attitudes to the use of healthcare data for research in England received a setback with the attempted implementation of the 'care.data' system that was intended to link general practitioner and hospital records at an individual patient level [5]. Poorly distributed information, the inability to opt out of the system readily, privacy concerns and the possible involvement of commercial bodies all contributed to a public outcry that prevented the scheme being implemented. Unfortunately, the mishandled implementation appears to have had a wider effect on public attitudes and professional responses to the use of healthcare data for research [4]. Even with the increased regulation recently introduced to ensure data security [35], there have been substantial delays in receiving data, and researchers should be aware that obtaining HES data may be difficult at the current time (MRC Health Economics Resource Use and Costs Working Group, personal communication).

5 Conclusions

In the context of a patient group of older men, the HES outpatient dataset appears valid for use in health economic evaluations conducted alongside RCTs or other study

designs, particularly following accreditation in 2008. Owing to the poorly completed diagnostic and procedure information, it is not currently suitable for use in economic evaluations specifically considering only condition-related resource use. The study was carried out within a patient group comprising men with prostate cancer. There is no reason to believe that the results are not generalisable to other patient groups, although this assumption needs to be tested; we have provided a straightforward methodology by which trial data could be used to test validity for additional patient groups in both the HES outpatient dataset and other routinely collected datasets in the UK and worldwide. The HES outpatient dataset offers many practical advantages, and some improvement in information, and may be a suitable alternative to collecting MR data manually within a trial.

Acknowledgments The authors acknowledge the contribution of the CAP trial group. The CAP trial is funded by Cancer Research UK/UK Department of Health (C11043/A4286, C18281/A8145, C18281/A11326 and C18281/A15064). This work was undertaken with the support of the MRC ConDuCT-II Hub (Collaboration and innovation for Difficult and Complex randomised controlled Trials In Invasive procedures—MR/K025643/1). Professors Jenny L. Donovan, Freddie C. Hamdy and David E. Neal are NIHR Senior Investigators.

Author contributions Joanna C. Thorn, Emma Turner, Eleanor Walsh, Richard M. Martin and Sian M. Noble were involved in the design, implementation and analysis of this methodological study. Jenny L. Donovan, David E. Neal, Freddie C. Hamdy and Richard M. Martin were involved in the design and implementation of the CAP trial on which the study is based. Eleanor Walsh, Julia Verne and Luke Hounsborne were involved in managing the data used in the study. All authors commented on drafts of the final manuscript.

Compliance with Ethical Standards

The CAP trial is funded by Cancer Research UK/UK Department of Health (C11043/A4286, C18281/A8145, C18281/A11326 and C18281/A15064). This work was undertaken with the support of the MRC ConDuCT-II Hub (Collaboration and innovation for Difficult and Complex randomised controlled Trials In Invasive procedures—MR/K025643/1). Professors Jenny L. Donovan, Freddie C. Hamdy and David E. Neal are NIHR Senior Investigators. Joanna C. Thorn, Emma Turner, Luke Hounsborne, Eleanor Walsh, Jenny L. Donovan, Julia Verne, David E. Neal, Freddie C. Hamdy, Richard M. Martin and Sian M. Noble have no conflicts of interest to declare.

References

1. Raghupathi W, Raghupathi V. Big data analytics in healthcare: promise and potential. *Health Inf Sci Syst.* 2014;2(1):3.
2. Lewsey J, Leyland A, Murray G, Boddy F. Using routine data to complement and enhance the results of randomised controlled trials. *Health Technol Assess.* 2000;4(22):1–45.
3. Williams J, Cheung W, Cohen D, Hutchings H, Longo M, Russell I. Can randomised trials rely on existing electronic data? A feasibility

- study to explore the value of routine data in health technology assessment. *Health Technol Assess.* 2003;7(26):1–117.
4. Border P. Big data and public health. Postnote 2014. <http://www.parliament.uk/business/publications/research/briefing-papers/POST-PN-474/big-data-and-public-health>. Accessed 30 Apr 2015.
 5. Health and Social Care Information Centre. NHS hospital data and datasets: a consultation. 2013. <http://www.england.nhs.uk/wp-content/uploads/2013/07/hosp-data-consult.pdf>. Accessed 24 Mar 2015.
 6. Health and Social Care Information Centre. Hospital Episode Statistics. <http://www.hscic.gov.uk/hes>. Accessed 30 Mar 2015.
 7. Health and Social Care Information Centre. Data dictionary: outpatients. http://www.hscic.gov.uk/media/1359/HES-Hospital-Episode-Statistics-HES-Outpatient-Data-Dictionary/pdf/HES_Outpatient_DD_Dec10.pdf. Accessed 22 July 2015.
 8. Health and Social Care Information Centre. Hospital outpatient activity—2013–14. 2015. <http://www.hscic.gov.uk/catalogue/PUB16722>. Accessed 30 Mar 2015.
 9. Health and Social Care Information Centre. Cleaning rules: outpatients. http://www.hscic.gov.uk/media/1368/HES-Hospital-Episode-Statistics-Outpatient-cleaning-rules/pdf/OPCR_July09.pdf. Accessed 30 Mar 2015.
 10. Williams J, Mann R. Hospital episode statistics: time for clinicians to get involved? *Clin Med.* 2002;2(1):34–7.
 11. Aylin P, Lees T, Baker S, Prytherch D, Ashley S. Descriptive study comparing routine hospital administrative data with the Vascular Society of Great Britain and Ireland's National Vascular Database. *Eur J Vasc Endovasc Surg.* 2007;33(4):461–5.
 12. Garout M, Tilney HS, Tekkis PP, Aylin P. Comparison of administrative data with the Association of Coloproctology of Great Britain and Ireland (ACPGBI) colorectal cancer database. *Int J Colorectal Dis.* 2008;23(2):155–63.
 13. Dixon J, Sanderson C, Elliott P, Walls P, Jones J, Petticrew M. Assessment of the reproducibility of clinical coding in routinely collected hospital activity data: a study in two hospitals. *J Public Health.* 1998;20(1):63–9.
 14. Burns EM, Rigby E, Mamidanna R, Bottle E, Aylin P, Ziprin P, Faiz OD. Systematic review of discharge coding accuracy. *J Public Health.* 2011;34(1):138.
 15. The Information Centre. Outpatient data quality report: 2003–04 and 2004–05. 2006. <http://www.hscic.gov.uk/catalogue/PUB02543/outp-qual.pdf>. Accessed 30 Mar 2015.
 16. Britton A, Milne B, Butler T, Sanchez-Galvez A, Shipley M, Rudd A, et al. Validating self-reported strokes in a longitudinal UK cohort study (Whitehall II): extracting information from hospital medical records versus the Hospital Episode Statistics database. *BMC Med Res Method.* 2012;12(1):83.
 17. Gaughan J, Mason A, Street A, Ward P. English hospitals can improve their use of resources: an analysis of costs and length of stay for ten treatments. CHE Research Paper 78. 2012.
 18. Thomas G. A cost-benefit analysis of the immunisation of children against respiratory syncytial virus (RSV) using the English Hospital Episode Statistics (HES) data set. *Eur J Health Econ.* 2015. doi:10.1007/s10198-014-0662-9.
 19. Bradshaw J, Noble M, Bloor K, Huby M, McLennan D, Rhodes D, et al. A child well-being index at a small area level in England. *Child Indic Res.* 2009;2(2):201–19.
 20. Marques E, Noble S, Blom AW, Hollingworth W. Disclosing total waiting times for joint replacement: evidence from the English NHS using linked HES data. *Health Econ.* 2014;23(7):806–20.
 21. Bloor K, Freemantle N, Maynard A. Gender and variation in activity rates of hospital consultants. *J R Soc Med.* 2008;101(1):27–33.
 22. Lambrelli D, Burge R, Raluy-Callado M, Chen S-Y, Wu N, Schoenfeld MJ. Retrospective database study to assess the economic impact of hip fracture in the United Kingdom. *J Med Econ.* 2014;17(11):817–25.
 23. Turner E, Metcalfe C, Donovan J, Noble S, Sterne J, Lane J, et al. Design and preliminary recruitment results of the Cluster randomised triAl of PSA testing for Prostate cancer (CAP). *Br J Cancer.* 2014;110(12):2829–36.
 24. Lane JA, Donovan JL, Davis M, Walsh E, Dedman D, Down L, et al. Active monitoring, radical prostatectomy, or radiotherapy for localised prostate cancer: study design and diagnostic and baseline results of the ProtecT randomised phase 3 trial. *Lancet Oncol.* 2014;15(10):1109–18.
 25. Department for Communities and Local Government. English indices of deprivation 2010: guidance document. 2011. <https://www.gov.uk/government/publications/english-indices-of-deprivation-2010-guidance>. Accessed 30 Mar 2015.
 26. Williams NJ, Hill EM, Ng SY, Martin RM, Metcalfe C, Donovan JL, et al. Standardisation of information submitted to an endpoint committee for cause of death assignment in a cancer screening trial—lessons learnt from CAP (Cluster randomised triAl of PSA testing for Prostate cancer). *BMC Med Res Method.* 2015;15(1):6.
 27. Health and Social Care Information Centre. Healthcare Resource Groups (HRG4). <http://www.hscic.gov.uk/hrg4/>. Accessed 30 Mar 2015.
 28. StataCorp. Stata Statistical Software: Release 12. College Station, TX: StataCorp LP; 2011.
 29. Benchimol EI, Manuel DG, To T, Griffiths AM, Rabeneck L, Guttmann A. Development and use of reporting guidelines for assessing the quality of validation studies of health administrative data. *J Clin Epidemiol.* 2011;64(8):821–9.
 30. Audit Commission. By definition: improving data definitions and their use by the NHS. 2011. <http://www.audit-commission.gov.uk/SiteCollectionDocuments/Downloads/20120419ByDefinition.pdf>. Accessed 11 Dec 2012.
 31. Head RF, Byrom A, Ellison GT. A qualitative exploration of the production of Hospital Episode Statistics in a Guernsey hospital: implications for regional comparisons of UK health data. *Health Serv Manag Res.* 2008;21(3):178–84.
 32. Thorn JC, Coast J, Cohen D, Hollingworth W, Knapp M, Noble SM, et al. Resource-use measurement based on patient recall: issues and challenges for economic evaluation. *Appl Health Econ Health Policy.* 2013;11(3):155–61.
 33. Spencer SA, Davies MP. Hospital Episode Statistics: improving the quality and value of hospital data: a national internet e-survey of hospital consultants. *BMJ Open.* 2012;2(6):e001651.
 34. Health and Social Care Information Centre. HES 2013–14 Annual Outpatient Data Quality Note. 2015. <http://www.hscic.gov.uk/catalogue/PUB16722/hosp-outp-acti-2013-14-op-qual-rep.pdf>. Accessed 30 Mar 2015.
 35. Health and Social Care Information Centre. Information Governance Toolkit. <https://www.igt.hscic.gov.uk/>. Accessed 31 Mar 2015.