Electronic primary dental care records in research: a case study of validation and quality assurance strategies

Kristina L Wanyonyi*[1, 2] David R Radford [1,3] Jennifer E Gallagher [2]

[1]University of Portsmouth Dental Academy, Hampshire Terrace, PO1 2QG, Portsmouth, UK

[2]King's College London Faculty of Dentistry, Oral & Craniofacial Sciences SE5 9RSLondon, UK

[3] King's College London Dental Institute, Teaching Division, Guys Tower, Guys Hospital, SE1 9RT, London, UK

Corresponding author*

Email: kristina.wanyonyi@port.ac.uk (KLW)

Highlights

1. 1. Dental records used in research can provide insights into task division, skill mix and needs-based workforce planning
2. 2. A stringent protocol of data cleaning and validation is required before electronic dental records are used
3. 3. Validated electronic dental records can be used to undertake research around patient risk factors
4. 4. Researchers need to work with clinicians and software developers to obtain rich and reliable data for dental research

Abstract

Background: In dentistry, the use of electronic patient records for research is underexplored. The aim of this paper is to describe a case study process of obtaining research data (sociodemographic, clinical and workforce) from electronic primary care dental records, and outlining data cleaning and validation strategies. This study was undertaken at the University of Portsmouth Dental Academy (UPDA), which is a centre of education, training and provision of state funded services (National Health Services). UPDA's electronic patient management system is R4/Clinical +. This is a widely used system in general dental practices in the UK. **Method:** A two-phase process, involving first Pilot and second Main data extraction were undertaken. Using System Query Language (SQL), data extracts containing variables related to patients' demography, socio-economic status and dental care received were generated. A data cleaning and validation exercise followed, using a combination of techniques including Maletic and Marcus's (2000) general framework for data cleaning and Rahm and Haido's (2010) principles of data cleaning. **Results:** The findings of the case study support the use of a two-phase data extraction process. The data validation processes highlighted the need for both manual and analytical strategies when cleaning these data. Finally, the process demonstrated that electronic dental records can be validated and used for epidemiological and heath service research. The potential to generalise findings is great due to the large number of records. There are, however, limitations to the data which need to be considered, relating to quality (data input), database structure and interpretation of data codes. **Conclusion:** Electronic dental records are useful in health service research, epidemiological studies and skill mix research. Researchers should work closely with, clinicians, managers and software developers to ensure that the data generated are accurate, valid and generalizable. Following data extraction, the research need to adapt stringent validation and data cleaning strategies to guarantee that the extracted electronic data are accurate.

Key words

Electronic dental records, Primary dental care, Dental informatics, Dental research, Clinical research, Electronic health records, Health service research

Background

Healthcare organisations generate a sizeable amount of data through health records. These data can inform our understanding of health services and patient management. In the past these data were in the form of paper records, and through enhanced informatics, they have developed into digitally stored records (1). This digitisation of health records, has created substantial data related to patients' medical history, care received, social circumstance and attendance patterns (2, 3). The initial drive for this development was to improve administrative functions of healthcare organisations (4), by documenting the needs and the care patients received over time, in order to facilitate communication between providers (1, 5). As accuracy and sophistication of these systems has improved, they have been developed to manage payments for care and planning clinical activity (6). Although comprehensive, these administrative functions only exploit the operational capabilities of these digital records. With time, the use of the analytical functions have emerged, and this includes decision support analytics, which provide clinicians with on-screen cues and prompts to guide their practice in order to improve both clinical outcomes and adherence to evidence-based guidelines (7).

In dentistry, the majority of patients are managed within a primary dental care setting. In England, in 2009, the Steele Review of 'NHS dental services in England' revealed that the majority (70%) of primary dental care practices submitted data to the NHS Business Service Authority (BSA) electronically as part of the payments system. The report further recommended 100% digitisation by 2011 (8). As expected, more widespread use of electronic systems has increased the quality and quantity of patient care data. In the near future, in England, there will be a standardised nomenclature for use in electronic patient records across primary care patient management systems using Systematized Nomenclature of Medicine-Clinical Terms (SNOMED CT) by 2019 (9).

With all the benefits of these real time electronic patient management systems, high volume, variety and velocity data are being generated, commonly referred to as ‘big data’ (10). Mining of these data for research, has the potential, to improve dental health surveillance and epidemiology (3), which at the moment is reliant on expensive surveys.

Electronic data have the potential to identify new patient-diagnostic groups and reveal unknown disease correlations (2). This could be undertaken through a process of predictive modelling, thereby influencing treatment decisions and disease progression through an informed change of the clinical pathway (11). Symptoms and diagnosis progression varies from patient to patient (12), and with the ever growing burden of chronic illnesses and an aging population with complex and seemingly unpredictable health outcomes, the ability to predict patterns in disease, and care that leads to better health outcomes is required.

In medicine, researchers have undertaken retrospective studies using the ever-growing repositories of observational data stored in electronic medical records (EMR). These studies have investigated guidance compliance, diseased patient identification (13) and general clinician practice (14). In dentistry across the UK, the limited research in the use of electronic records, has involved the analysis of dental treatment data to ascertain the longevity of treatment materials, and time to re-intervention after treatment (15-21), and more recently even time to tooth extraction (22-26). There has been limitated use of these data for epidemiological research as the information collected centrally within the NHS has related to payment and focused on treatments; and has not been sufficiently granular for epidemiological research. This has also been because primary care dentistry has yet to develop a commonly accepted standardised terminology to describe oral diagnoses,; lagging behind medicine in its codification of diagnoses (27).

The work within our current study, represents a detailed approach of extracting more granular data, when compared to the mentioned previous studies and augmenting these data to ascertain the profile of patient disease and risk of disease. From one large dental practice, the

data generated answered questions related to individual patient experiences which resulted in four publications. These detailed the relationship between dental access, geography and socio-demography, how skill mix occurs in general dental practice, equity in treatment provision by patient social circumstance and modelled alternative scenarios for preventive care (28-31). Dental researchers have the potential to explore these data to undertake more predictive studies to identify diagnostic patient groups, ideal care pathways and risk factors for poor outcomes (30). During the process of our study, it was clear that data quality remains a poorly researched issue, particularly in the field of dentistry. There are no agreed data quality assessment framework to undertake data quality assessment in electronic health records (32); however, a general consensus around data accuracy (33), completeness (34), consistency, credibility, and timeliness has been agreed as key (32). There is also no guidance on how to deal with any data quality issues to gain a research usable data set. Therefore, through this work, additional insights have been gained on how to validate, clean and use these data for research.

The aim of this paper is to describe the process of obtaining research data from electronic primary care dental records, outlining data quality assessment and validation strategies, followed by data cleaning, which consist of dealing with missing data, determining record usability, and identifying erroneous data, which link to the key areas of data quality assessment accuracy (34), completeness (35), consistency, credibility, and timeliness. We drew on two approaches in our data cleaning and validation process: principles from Rahm and Hai Do, [35], and Maletic and Marcus [36) . These approaches involved i] defining and determining error types, ii] searching and defining error instances iii] correcting the uncovered error instances.

Methods

The research was undertaken using data extracted from the University of Portsmouth Dental Academy (UPDA). UPDA is a state funded National Health Service (NHS) primary dental care service provider and undergraduate training centre for dental professionals. Ethical approval for this research was provided by NRES Committee Fulham REC: Reference No. 11/LO/1138 Protocol No. NTMHWMOV3 and research governance approval by NHS Portsmouth R&D Committee Reference No. SSPS/05/11. Patients attending this facility were able to identify if they did not wish their data to be part of research and posters in the health centre provided further information in line with Caldicott guidelines. UPDA uses a live electronic patient management system, which collects all relevant patient and clinical data, only some of which submits data to the National Health Service Business Authority (NHS BSA) in order to fulfil the existing contract. The data generated from these systems were therefore considered valid accounts used for remuneration for services. An exercise to extract and clean the electronically stored data for use in this research was then undertaken. The aim was to obtain a dataset that included patient characteristics, dental care provided and the nature of the care provider.

The development of the project protocol involved team discussions with software developers (Carestream Ltd), clinicians who input the data, and social science researchers. As big data analysis is question driven, the study was informed by the literature on inequalities in oral health, access to dental care and the potential for use of skill mix in meeting the growing demands for dental care. The extraction of data followed a thorough system appraisal, consultation with software developers using a database schema for the R4/Clinical + software, and writing of a data extraction script in Structured Query Language (SQL). This is described in Figure 1 with further details of the SQL in Supplementary Files 1 and 2. Data were held in real time, thus the pilot and main dataset accorded with different time periods.

The approach to handling the data was as follows

I] Defining and determining error types

In order to effectively define the error types a two-phase process was undertaken which included a pilot extract which would inform a main extract. Using the pilot extract data Rahm and Hai Do, 2013, and Maletic and Marcus (2000), techniques of defining errors through interrogation was undertaken. This involved descriptive analysis to establish completeness and inaccuracies in the data underpopulated or missing data, mismatch in names and errors of duplication of data. Frequency statistics were used, as defined in Feder's systematic review on data quality assessment in use of electronic health records (32). The next step was to uncover the sources of errors.

ii] Searching and defining error instances

The errors were classified as to whether they were at the instance or the schema. The schema can be described as a "layout" of a database or the blueprint that outlines the way data are organised into tables (33). Schema-level problems are reflected in the instances; they can be addressed at the schema level by an improved schema design (schema evolution), schema translation and schema integration (34). In this case, this was the type of codes within the system or the way data was named on the coded-system. Instance-level problems, on the other hand, refer to errors and inconsistencies in the actual data contents, which are not visible at the schema level (34). This could be how data was input by clinicians. The Pilot extract was useful in identifying schema errors/problems which limited the use of certain variables. The instance-level errors were the primary focus of data cleaning in in the main extract.

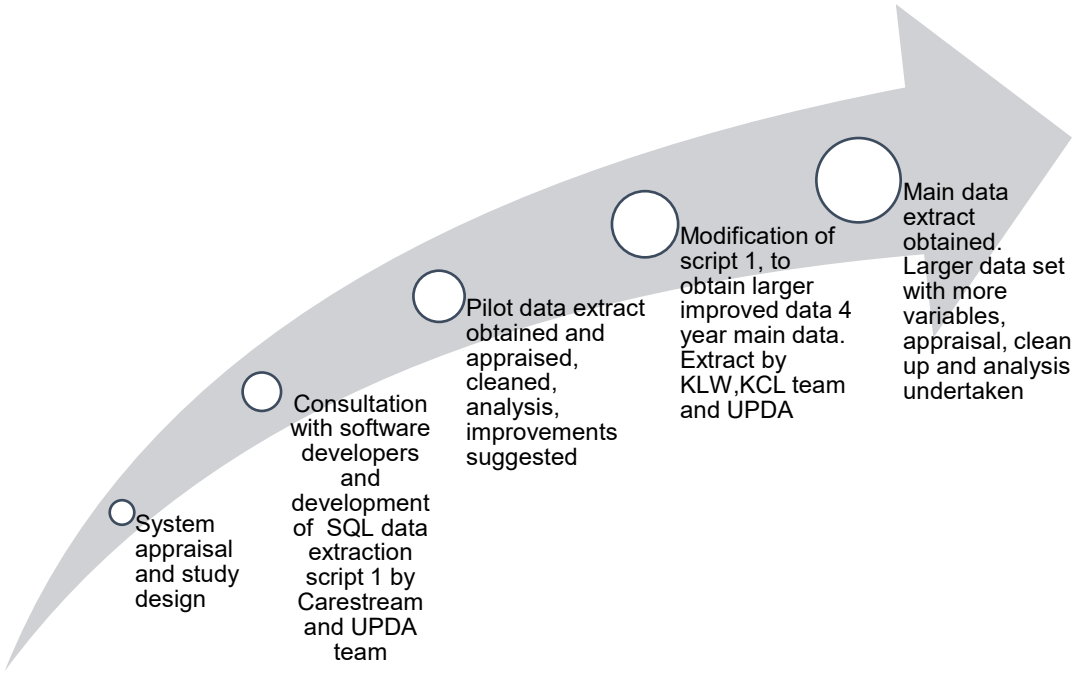
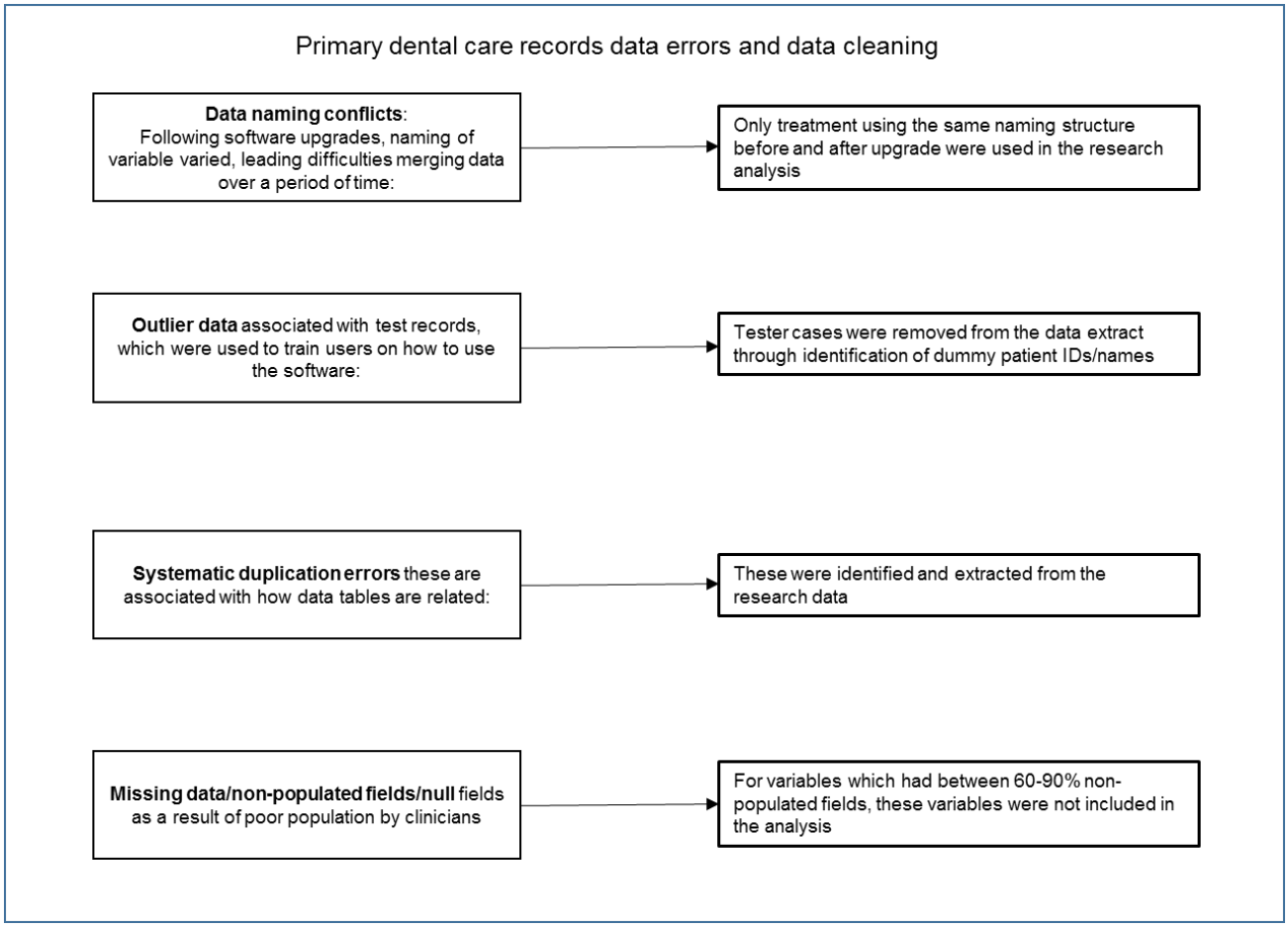


Figure 1: Methodology

iii] Correcting the uncovered error instances

Reflecting on Maletic and Marcus's (2000), general framework for data cleaning, the pilot extract allowed us to identify the errors likely to be encountered in the main extract (Fig 3). As the use of electronic patient management data for dental research is fairly new, the use of a combination of techniques for validation (35-39), was necessary in order to gain a robust data set. The information technology cleaning strategies (35, 38), aided the process, and the health service studies (36, 37, 39), facilitated the development of further manual validation.



Adapted from Rahm and Hai Do, 2013 and Maletic and Marcus, 2000

Figure2 Big Data errors and data cleaning of primary dental care electronic records

Manual data validation/onsite validation was undertaken for the pilot extract and main extract. This method has been used by Hall et al. (2008), to evaluate aggregated data bases from a variety of clinics, and has been shown to provide successful insight into validity of retrospectively extracted data. Thomas et al. (2014) prospectively collected the data manually then compared this to the outputs from the electronic system and found that there was accuracy in the outputs of the electronic data.

Results

The result was two datasets (pilot and main data) were extracted in this project, the former informing the latter. Both datasets were cleaned to allow appropriate data analysis as outlined below.

Pilot and main data extract charecteristics

The pilot extract included 4,343 patient records. These data comprised the last completed course of care for patients treated at UPDA between 1 September 2009 and 31 August 2011 (two academic years). The categories of patient variables (n= 10) were: Date of Birth, Sex, Postcodes, Ethnicity, Date of treatment, Benefit status, Oral health risks status (RAG rating), Treatment plan, Procedures undertaken (some including operators). Identifiable data such as Date of Birth and Post code were pseudo anonymised to age and Lower Super Output Area (LSOA) respectively. LSOA was subsequently converted to indices of multiple deprivation (IMD). LSOA represents 1,500 households in a geographic area in England and is used in census data in the UK to mark out an area of IMD(40). The pilot data informed a more robust main extract of 6,351 patients (Fig 2). The pilot data highlighted variables that were not populated and could not be used in the analysis. Lack of linkages in the relational tables based on the schema, limited some types of analysis e.g. band of course of care linking to patients. Thus the main extract was broader and was targeted. It consisted of all courses of care over a four-year period (2008 -2012). It had more variables than the pilot extract. Of particular importance in the main extract, was to establish a way to describe patient disease risk, as the risk variable (RAG score) in the pilot extract was poorly populated (10%). Smoking status was also not well populated in the pilot data, based on the schema field that was used. A proxy variable smoking cessation signposting which was held in the administrative fields of the schema had a 95% rate of identifying and signposting smokers to cessation services was therefore included in the main extract. The number of patients in the main extract remained the same after cleaning. The cleaning involved deduplication of observations and clearing-naming conflicts, but the data was from the same number of patients.

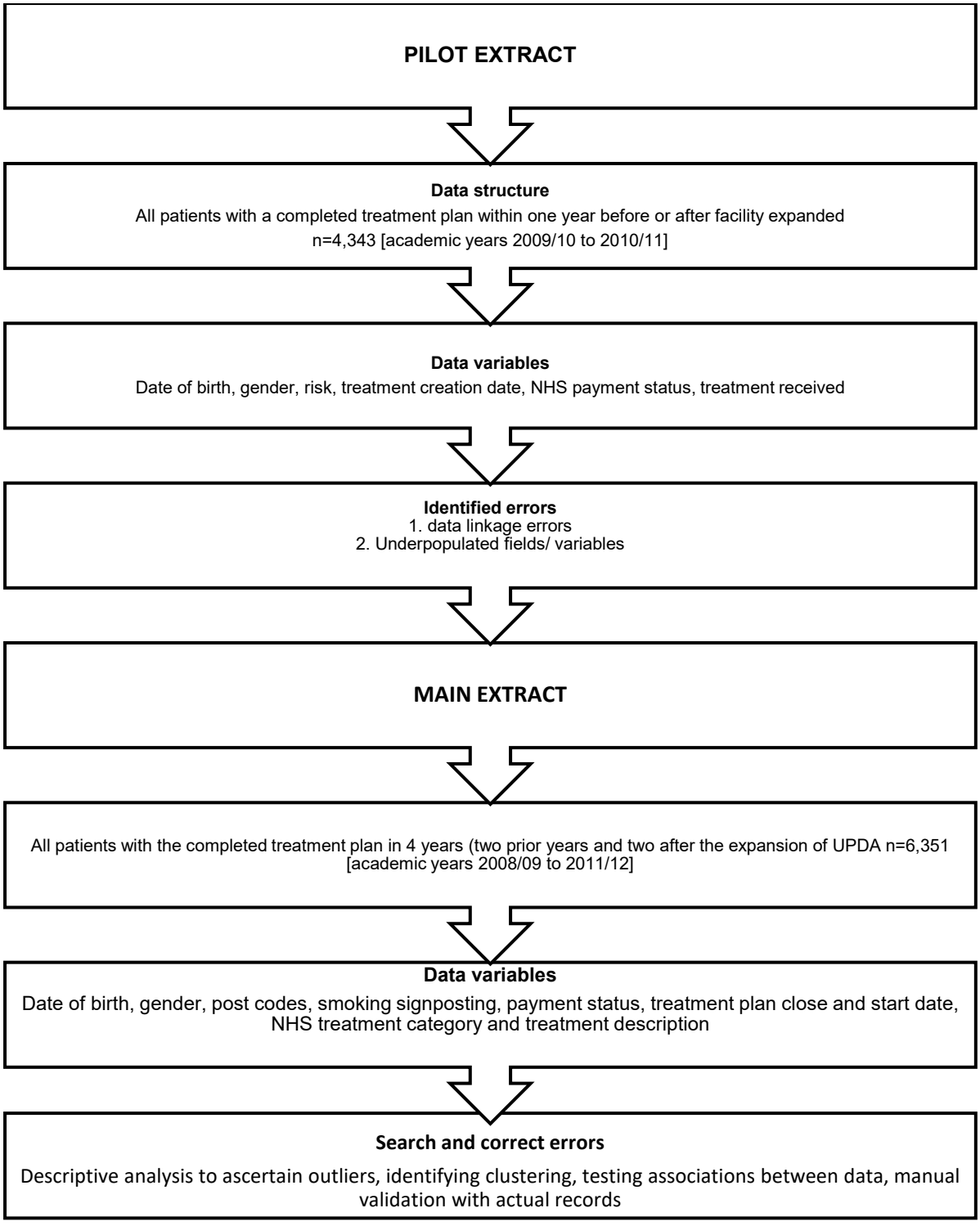


Figure 3 Big data extraction and analysis phases

I] The sources of errors in data

There were both schema and instance level errors in these data. The instance-level errors were the primary focus of data cleaning in post extraction phase; this type of error is associated with how the users of the software input data. Figure 3 shows errors in UPDA's data and the solutions employed in cleaning and validation. These quality problems were; naming conflicts, poor schema design, misspellings and duplicates.

ii] defining the errors

Through the pilot data it was highlighted that certain variables were not fully populated and could not be used in the analysis and it showed where there was a lack of data linkages in the relational tables based on the schema which limited some types of analysis e.g. band of course of care linking to patients. As a whole the pilot data showed what we could reliably gain from the data and how we could amend the data extraction script in the second phase in order to gain more data and to select different variables in some cases. The main extract was, therefore, broader and was targeted. It consisted of all courses of care over a four-year period (2008 -2012). It had more variables than the pilot extract. Of particular importance in the main extract, was to establish a way to describe patient disease risk, as the risk variable (RAG score) in the pilot extract was poorly populated (10%). This score is a score which is used to determine a pathway of care. It is not possible for a course of care to be closed without an assessment of RAG score. It was concluded that this was not well populated using the coded system. It was later established that this variable could have been placed on the system in pictorial format, hence giving clinicians two options on how to record it. Smoking status was also not well populated in the coded variable we extracted in the pilot data and we established that this was underpopulated because it was a key performance indicator (therefore recorded in a different domain?????) and all patients are asked this question. As a proxy variable 'smoking cessation signposting' which had a 95% rate of identifying and signposting smokers to cessation services as highlighted in the NHS BSA reports provided to UPDA. This variable was therefore included in the main extract.

Iii] Correcting the uncovered error instances

Maletic and Marcus's (2000), general framework for data cleaning, involved statistical analysis of outliers, identifying clustering, testing associations between data, and manual validation against actual clinical records for a sample of cases. Equally, techniques by Feder 2018 (32) for general medical electronic records also suggest similar techniques and triangulation techniques.

Manual data validation/onsite validation was undertaken for the pilot extract and main extract. The process in this research involved scanning for 50 patient IDs from already extracted data. Using the user interface on-site at the clinic, these records were retrieved and actual patient records were obtained and checked against the extracted treatment plan data and patient details. The results showed matching of patient details in the data set with patient records. There was further validation with external data sets. Data from the NHS BSA. authority was analysed against the study extract data. The age profile and treatment by age variables were analysed and compared to data reports from the NHS BSA. Other studies have used self-reports to manually triangulate or test the accuracy of electronic records in a similar way (33).

Research Outputs

This project resulted in four peer reviewed publications (28-31). The first paper from the pilot dataset was limited to two years and allowed the team to research the relationship between access to UPDA before and after expansion with patients' demography and deprivation (28). Although the data were cross-sectional, being able to identify patients seen in two distinct

periods based on completed treatments, it was possible to uncover patterns of inequality in access to dental services that were solely related to deprivation. The second study highlighted the distribution of task between dental and dental care professionals (DCPs) in training (29). These data were the first of its kind, as national data sets do not document whether care was undertaken by dentists or DCP, but care provided by a general dental practice. These data were further used to apply to an operational research model, in the third study, which tested a variety of scenarios where skill mix was used to increase preventative care cost-effectively (31). The fourth study, investigated the predictors of dental treatment. Due to the large data and ability to augment the data on patient residence to deprivation status, it was possible to investigate whether deprivation and area of residence predicted the receipt of advanced care, and/or preventive care. In the future, exploring how dental data can be linked to general medical practice or hospital data is an area in need of analysis.

Discussion

This paper describes a case study process in which electronic dental records are extracted, validated and use in researchs. The process identifies a framework that produced reliable research data set, as evidenced by validation of the data against patterns of treatment from national data and local administrative reports. Primary dental care records hold a wealth of information on patients' demography, care received, and the provider. If augmented with other national data sets, such as residential deprivation data and census data, it is possible to undertake predictive analysis on patients who are at risk for disease or who might have varied dental care needs.

Data cleaning is a major part of the process and a combination of techniques from informational technology (35-39) were found to be reliable in obtaining clean data. Identifying errors in the data storage and relational tables (schema) and how the data are inputted (instance) was fundamental to the data cleaning process. These identified errors included

naming conflicts, outlier data, duplications and missing data, all of which were identified in two-stages.

As previously highlighted, augmenting electronic patient data with other administrative data such as census records to obtain more information related to patients was instrumental in answering the research questions related to identifying individual and societal factors that which contribute to disease risk. For this exercise, post codes were used to obtain contextual information relating to where a patient lived. The indices of multiple deprivation is an index made up of several domains which describe how deprived an area of residence is in relation to factors such as income, health disability and access to health services (40). It is important to consider ethical standards related to confidentiality when using post code data. Researchers need to ensure, that the data are converted to the target (deprivation score) as soon as possible variable and any identifiable versions of the data are destroyed in compliance with data protection laws.

Text data mining, which is now possible within general medicine research (41), is an area for future exploration in dentistry. Challenges exist in obtaining accurate non-codable data from free text as natural language processing has not developed to a point where coded data can be replaced (42, 43). It may be possible to filter some string text, but non-coded data is not normally standardized in dentistry, and it is important to ensure that the string-format text is interpreted accurately (44). It has been proposed that in order to improve non-coded data, guidelines should be put in place relating to words and texts for clinicians and those inputting information into the system (45). However, it is important to recognise that clinicians would vary in their description of certain aspects of care and dentistry has a limited number of diagnostic codes. On occasion, mining all text may be an option, however, there is the risk of extracting identifiable information including names, e.g ‘Mr Smith has been advised to brush his teeth’. In dentistry, more discussions on free text capture will be necessary. However, at present, the use of systems such as SNOMED –CT in dentistry (9), will allow conversion of some previous free text information into a coded accessible structure.

Accurate data entry remains the key to extracting a robust and high quality data set for research purposes. A clear policy for clinicians relating to input of data is helpful and should ideally be conducted in dialogue with researchers to achieve the desired outcomes. Public and patient participation should also be considered in line with contemporary research practice (46) to inform the type of data collected within patient management systems. Other ways of ensuring that data fields are populated accurately is to ensure that where defaults can be applied these are used. For example, if a patient is aged under 18 years, a default within the payment field would show that patient is exempt from payment as they are a child. As a routine data quality approach, where a variable of interest is poorly populated, omitting this from any analysis ensures no biases are introduced to the data (28-30). In this study, it was possible to substitute the variables as there was a pilot phase. Finding out that often the schema collects data in multiple fields was a useful exercise. Additionally, ensuring providers of care are coded into the data ensures that more research into the use of skill mix can be undertaken (29).

Other limitations, beyond the researcher's control, relate to the data schema. The schema is the map of the data storage tables (47). This is dependent on the software developers. Some of the challenges with the schema could include missing information, or changes in the way data are stored due to system upgrades. Finally, these data are often cross-sectional and the nature of databases changes a lot and data are regularly overridden. This is very serious as it limits the potential for longitudinal research on certain data to achieve a deeper understanding of disease, treatment processes and outcomes. An approach to overcome this is to create data warehouses which collect and retain episodes of information, allowing the longitudinal information to remain stored and be linkable using the unique patient identifier (NHS number). Such a facility should also ideally enable the linking of primary medical and dental care to be linked to one another and to secondary care.

Conclusions

The primary users' data entry processes need to be streamlined to ensure appropriate population of data. Software developers need to carefully align syntax and variables following system upgrades to previous versions of software in order to ensure homogeneity of variable names across different periods. Researchers need to adapt stringent validation and cleaning strategies to guarantee that the electronic data used in research are accurate. Having a data warehouse structured to collect data periodically in a format that records episodes and enables data linkages could enhance research in this field.

Summary points

What was already known

- Electronic dental records can provide insights into the quality of dental care provided particularly longevity of restorations
- In the England, there is a widespread drive to increase the use of electronic dental records for the monitoring the provision of state-funded care
- Although the national data sets on demand and provision of state-funded dental care have been made available, no individual level data analysis of dental care received has been undertaken

What this study has added

- This is the first study first to explore the extraction and use of individual level electronic dental records in research of state-funded care, in order to explore social factors, skill mix and other predictors of dental needs
- This is the first study to describe a process of mining electronic dental data at the primary dental care level where the majority of care is received, and to highlight ways to ensure good data quality

- This paper provides useful insights on how to reliably expand on the use of electronic dental data for further dental research.

Funding

This research was funded through a PhD studentship supported by King's College London and the University of Portsmouth Dental Academy.

Acknowledgements

We acknowledge the support of the University of Portsmouth Dental Academy and Carestream Ltd.

Conflict of interest

DRR was Director of Clinical Studies for the Dental Students at UPDA, whilst JEG leads Dental Public Health teaching across KCLDI. KLW contributed to the undergraduate dental teaching programme at KCLDI and is now on the staff of UPDA.

References

1. I 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):470–477.
2. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. Nat Rev Genet. 2012 06//print;13(6):395-405.
3. Birkhead GS, Klompas M, Shah NR. Uses of electronic health records for public health surveillance to advance public health. Annual review of public health. 2015 Mar 18;36:345-59. PubMed PMID: 25581157. Epub 2015/01/13. eng.
4. Anderson J. Clearing the way for physicians' use fo clinical information systems. Commununications of the ACM. 1997;40(8):83-90.
5. Eggleston E, Klompas M. Rational Use of Electronic Health Records for Diabetes Population Management. Current Diabetes Reports. 2014 2014/03/11;14(4):1-10. English.
6. Bose R. Knowledge management-enabled health care management systems: capabilities, infrastructure, and decision-support. Expert Systems with Applications. 2003;24(1):59-71.
7. Vikram K, Karjodkar FR. Decision Support Systems in Dental Decision Making: An Introduction. Journal of Evidence Based Dental Practice. 2009;9(2):73-6.
8. Steele J. A review of NHS Dental Services in England: an independent review2009 04/01/2019. Available from: http://www.sigwales.org/wp-content/uploads/dh_101180.pdf.
9. NHS Digital. SNOMED CT2018. Available from: https://digital.nhs.uk/services/terminology-and-classifications/snomed-ct.
10. Heller KE, Eklund SA, Burt BA, Briskie DM, Lawrence LM. Using Insurance Claims and Demographic Data for Surveillance of Children's Oral Health. Journal of Public Health Dentistry. 2004;64(1):5-13.
11. Ng K, Ghoting A, Steinhubl SR, Stewart WF, Malin B, Sun J. PARAMO: A PARAllel predictive MOdeling platform for healthcare analytic research using electronic health records. Journal of Biomedical Informatics. 2014 2014/04/01/;48:160-70.
12. Kidd E. Essentials of Dental Caries: The disease and its management. London: Oxford; 2005.
13. Hogan WR, Wagner MM. Accuracy of data in computer-based patient records. Journal of American Medical Informion Associaction. 1997;4(5):342-55.
14. Ashworth M, Cox K, Latinovic R, Charlton J, Gulliford M, Rowlands G. Why has antibiotic prescribing for respiratory illness declined in primary care? A longitudinal study using the General Practice Research Database. Journal of Public Health. 2004 September 1, 2004;26(3):268-74.
15. Burke FJT, Lucarotti PSK. Ten-year outcome of crowns placed within the General Dental Services in England and Wales. Journal of Dentistry. 2009 1//;37(1):12-24.
16. Burke FJT, Lucarotti PSK. Ten-year outcome of porcelain laminate veneers placed within the general dental services in England and Wales. Journal of Dentistry. 2009 1//;37(1):31-8.
17. Burke FJT, Lucarotti PSK. Re-intervention in glass ionomer restorations: What comes next? Journal of Dentistry. 2009 1//;37(1):39-43.
18. Burke FJT, Lucarotti PSK. Re-intervention on crowns: What comes next? Journal of Dentistry. 2009 1//;37(1):25-30.
19. Burke FJT, Lucarotti PSK, Holder RL. Outcome of direct restorations placed within the general dental services in England and Wales (Part 2): Variation by patients' characteristics. Journal of Dentistry. 2005 11//;33(10):817-26.
20. Lucarotti PSK, Holder RL, Burke FJT. Analysis of an administrative database of half a million restorations over 11 years. Journal of Dentistry. 2005 11//;33(10):791-803.
21. Lucarotti PSK, Holder RL, Burke FJT. Outcome of direct restorations placed within the general dental services in England and Wales (Part 3): Variation by dentist factors. Journal of Dentistry. 2005 11//;33(10):827-35.

22. Burke FJT, Lucarotti PSK. The ultimate guide to restoration longevity in England and Wales. Part 5: crowns: time to next intervention and to extraction of the restored tooth. Br Dent J. 2018 Jul 13;225(1):33-48. PubMed PMID: 29977023. Epub 2018/07/07. eng.
23. Burke FJT, Lucarotti PSK. The ultimate guide to restoration longevity in England and Wales. Part 4: resin composite restorations: time to next intervention and to extraction of the restored tooth. Br Dent J. 2018 Jun 22;224(12):945-56. PubMed PMID: 29999041. Epub 2018/07/13. eng.
24. Burke FJT, Lucarotti PSK. The ultimate guide to restoration longevity in England and Wales. Part 3: Glass ionomer restorations - time to next intervention and to extraction of the restored tooth. Br Dent J. 2018 Jun 8;224(11):865-74. PubMed PMID: 29855590. Epub 2018/06/02. eng.
25. Burke FJT, Lucarotti PSK. The ultimate guide to restoration longevity in England and Wales. Part 2: Amalgam restorations - time to next intervention and to extraction of the restored tooth. Br Dent J. 2018 May 25;224(10):789-800. PubMed PMID: 29795518. Epub 2018/05/26. eng.
26. Lucarotti PSK, Burke FJT. The ultimate guide to restoration longevity in England and Wales. Part 1: methodology. Br Dent J. 2018 May 11;224(9):709-16. PubMed PMID: 29747178. Epub 2018/05/11. eng.
27. Kalenderian E, Ramoni RL, White JM, Schoonheim-Klein ME, Stark PC, Kimmes NS, et al. The development of a dental diagnostic terminology. Journal of dental education. 2011;75(1):68-76. PubMed PMID: 21205730.
28. Wanyonyi KL, Radford DR, Gallagher JE. The relationship between access to and use of dental services following expansion of a primary care service to embrace dental team training. Public health. 2013 Nov;127(11):1028-33. PubMed PMID: 24210166. Epub 2013/11/12. eng.
29. Wanyonyi KL, Radford DR, Gallagher JE. Dental skill mix: a cross-sectional analysis of delegation practices between dental and dental hygiene-therapy students involved in team training in the South of England. Human Resources for Health. 2014;12(1):1-8.
30. Wanyonyi KL, Radford DR, Gallagher JE. Dental Treatment in a State-Funded Primary Dental Care Facility: Contextual and Individual Predictors of Treatment Need? PLoS ONE. 2017 01/24

04/08/received

12/09/accepted;12(1):e0169004. PubMed PMID: PMC5261606.
31. Wanyonyi KL, Radford DR, Harper PR, Gallagher JE. Alternative scenarios: harnessing mid-level providers and evidence-based practice in primary dental care in England through operational research. Human Resources for Health. 2015;13(1):1-12.
32. Feder SL. Data Quality in Electronic Health Records Research: Quality Domains and Assessment Methods. Western Journal of Nursing Research. 2018;40(5):753-66. PubMed PMID: 28322657.
33. Cole AM, Pflugeisen B, Schwartz MR, Miller SC. Cross sectional study to assess the accuracy of electronic health record data to identify patients in need of lung cancer screening. BMC Research Notes. 2018 January 10;11(1):14.
34. Alwhaibi M, Balkhi B, Alshammari TM, AlQahtani N, Mahmoud MA, Almetwazi M, et al. Measuring the quality and completeness of medication-related information derived from hospital electronic health records database. Saudi Pharmaceutical Journal. 2019 2019/01/17/.
35. Maletic JI, Marcus A. Data Cleansing: Beyond Integrity Analysis Information Quality (IQ2000) [Internet]. 2000 15/07/2013:[200-9 pp.]. Available from: http://dc-pubs.dbs.uni-leipzig.de/files/Maletic2000DataCleansingBeyond.pdf.
36. Hall GC, Bryant TN, Merrett LK, Price C. Validation of the quality of The National Pain Database for pain management services in the United Kingdom. Anaesthesia. 2008;63(11):1217-21.
37. Kudyakov R, Bowen J, Ewen E, West SL, Daoud Y, Fleming N, et al. Electronic health record use to classify patients with newly diagnosed versus preexisting type 2 diabetes:

infrastructure for comparative effectiveness research and population health management. Popul Health Manag. 2012;15(1):3-11.
38. Rahm E, Hai Do H. Data cleaning: problems and current approaches2013 17/10/2013. Available from: http://wwwiti.cs.uni-magdeburg.de/iti_db/lehre/dw/paper/data_cleaning.pdf.
39. Thomas A, Zheng C, Jung H, Chang A, Kim B, Gelfond J, et al. Extracting data from electronic medical records: validation of a natural language processing program to assess prostate biopsy results. World J Urol. 2014 2014/02/01;32(1):99-103. English.
40. NHS The Information Centre for Health and Social care. The Indices of Multiple Deprivation. 2012.
41. Raja U, Mitchell T, Day T, Hardin JM. Text mining in healthcare. Applications and opportunities. Journal of healthcare information management : JHIM. 2008 Summer;22(3):52-6. PubMed PMID: 19267032. Epub 2009/03/10. eng.
42. Chapman WW, Christensen LM, Wagner MM, Haug PJ, Ivanov O, Dowling JN, et al. Classifying free-text triage chief complaints into syndromic categories with natural language processing. Artificial Intelligence in Medicine. 2005;33(1):31-40.
43. de Lusignan S, van Weel C. The use of routinely collected computer data for research in primary care: opportunities and challenges. Family practice. 2006 April 2006;23(2):253-63.
44. Scobie S, Basnett I, McCartney P. Can general practice data be used for needs assessment and health care planning in an inner-London district? Journal of Public Health. 1995 December 1, 1995;17(4):475-83.
45. Van Weel-Baumgarten EM, Van den Bosch WJ, Van den Hoogen HJ, Zitman FG. The validity of the diagnosis of depression in general practice: is using criteria for diagnosis as a routine the answer? British Journal of General Practice. 2000 Apr;50(453):284-7. PubMed PMID: ISI:000086252800005. English.
46. Coiera E. Building a National Health IT System from the middle out. J Am Med Inform Assoc. 2009 //;16:271-3.
47. Chapple M. About databases: Schema2014 12/02/2012. Available from: http://databases.about.com/cs/specificproducts/g/schema.htm.