Streams, rivers and data lakes: an introduction to understanding modern electronic healthcare records

Authors: Esther Ayobamidele Abisola Idowu, A James Teo, B Sabrine Salih, C Joshua Valverde D and Joshua Au Yeung E

As foundation doctors, we have often found ourselves informing patients that a certain aspect of their medical information cannot be immediately found, either because it is on an electronic system we cannot access, or it is in a hospital that is unlinked to our own. Unsurprisingly, this frequently leaves patients flabbergasted and confused. We started to wonder: if patients’ data are entered onto an electronic system: where do those data go? If medical data are searched for, where do those data come from? Why are there so many hidden sources of information that clinicians cannot access? In an ever-increasing digital sphere, electronic data will be the future of holistic health and social care planning, impacting every clinician’s day-to-day role. From electronic healthcare records to the use of artificial intelligence solutions, this article will serve as an introduction to how data flows in modern healthcare systems.

KEYWORDS: electronic health records, healthcare data, artificial intelligence, data standardisation, data ethics

DOI: 10.7861/clinmed.2022-0325

What makes electronic healthcare records differ from one another?

Healthcare data can be defined as patient information that is entered by clinical or non-clinical individuals. Traditionally written in paper format (eg discharge letters, test results, appointments and clinical notes), patient records could become multiple volumes of healthcare data. Electronic healthcare records (EHR) are defined as a data system that stores patient healthcare data. Electronic healthcare records (EHR) are defined as a data system that stores patient healthcare data. Although a clinician might believe that they are interacting with a single EHR, in reality, most hospitals use multiple types (Table 1).

EHR software brands can vary by hospital and are often divided into two categories: (1) bespoke (‘lots of little’ EHRs or ‘best of breed’) models: where many single systems are designed for specific departments needs and data communicates between them; or (2) monolithic models: where one EHR system is used for the whole hospital serving every clinical department or scenario.

The benefit of monolithic EHRs is that they directly link healthcare data without the need to open new software or use multiple single sign-on tools. However, monolithic EHRs can be difficult to specialise to the needs of every clinical department. For foundation doctors, this manifests as a clear difference in usability from one EHR system to another.

Data fragmentation and integration: where are my patients’ data?

EHRs that are entirely unlinked to other healthcare software are termed ‘data silos’. Data silos represent a closed-off, unintegrated source of clinical information and include data used in research, private healthcare, genomics and in wearable devices (eg digital watches that monitor heart rate). Historically, when EHRs were deployed, not enough consideration was put into how those data could be shared or integrated into other systems. From a foundation doctor’s perspective, this means fragmented or missing data, which require accessing multiple systems to retrieve (eg by making laborious calls and sending emails to clinical secretaries or on-call doctors at another institute or NHS Trust).

Another way in which healthcare data become fragmented and lost is when a clinician attempts to adapt an EHR to fit their clinical need. This occurs via the frontend–backend disconnect: what a clinician sees on the user interface (frontend) does not encompass all the data stored (backend). It is relatively easy for clinicians to alter the frontend with new electronic forms. However, without altering the backend, these forms are stored in an unusable form; moreover, these alterations require specialist knowledge often lacking in NHS Trusts. Therefore, healthcare records mirror a library without an indexing catalogue, where anyone can just throw a book inside (Fig 1).

Preventing this ‘library-with-no-index’ scenario, where data cannot be easily retrieved, requires data linkage and integration. This has been attempted for more than 20 years and is currently happening via the integrated care systems (ICS), outlined in the NHS 2019 Long-Term Plan. Unfortunately technological barriers exist that prevent data linkage and integration, preventing clinicians from accessing healthcare data, thereby worsening the socioeconomic burden on healthcare systems.
Semantic standardisation is traditionally achieved through clinical coding. Clinical coding is performed by GPs in GP practices, and by clinical coders in hospitals. Clinical coders read and interpret clinical text to assign relevant codes that reflect a patient’s care during a hospital visit. Correct and efficient clinical coding is essential for service evaluation, research and medical billing.

Inaccuracies in coding (e.g. if a procedure that was carried out is missed) lead to the misallocation of staff, funds and resources, resulting in inefficiencies and waste. For foundation doctors, this often means the frustrating task of filling in structured proformas or checklists to help inform clinical coders before an administrative task is complete or a patient is discharged. The major coding system currently used for diseases is the International Classification of Diseases version 10 (ICD-10). Office of Population and Census survey version 4 (OPCS-4) is the coding classification used for interventions and procedures.

Another newer ontological coding system is called Systematized Nomenclature of Medicine-Clinical Terms (SNOMED-CT): this covers diseases, procedures as well as many other biomedical and healthcare concepts (e.g. anatomy, occupation, etc.). Unlike other

### Table 1. Examples of electronic healthcare records (EHRs), also known as electronic healthcare modules, with their uses

<table>
<thead>
<tr>
<th>EHR/module</th>
<th>Use</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient administration system (PAS)</td>
<td>Registration of patients, receipt of referrals, scheduling appointments and managing of hospital bed loads; often associated with a bed management module that looks at the state of the beds and allows for assessment of bed state</td>
</tr>
<tr>
<td>Infection management EHR</td>
<td>Analysis of the spread of infections</td>
</tr>
<tr>
<td>Labs EHR (pathology virology and microbiology), also known as a laboratory information system (LIM)</td>
<td>Management of the flow of samples through the laboratory workflow; samples are analysed and results uploaded</td>
</tr>
<tr>
<td>Radiology picture archive and communication system (PACS)</td>
<td>Vendor-neutral archive systems which store the images and the clinical information on the scan</td>
</tr>
<tr>
<td>Radiology information system (RIS)</td>
<td>Scheduling of radiology appointments and radiographer/sonographer workflow, recording type of scan and dosing parameters; link to radiology PACS</td>
</tr>
<tr>
<td>Community EHR</td>
<td>Out of hospital scheduling, offline access to records, recording of patient information during home visits</td>
</tr>
<tr>
<td>Theatre systems</td>
<td>Schedule patients, record theatre notes and allows for monitoring of theatre equipment</td>
</tr>
<tr>
<td>Emergency department system</td>
<td>Registering patients; often there are links to clinical modules to allow requesting</td>
</tr>
<tr>
<td>Order communication (order comms)</td>
<td>Often used for ordering diagnostic testing, e.g. for blood tests that require labelling and printing; often linked to the results module</td>
</tr>
<tr>
<td>Results EHR</td>
<td>Overview of results of diagnostic tests, which may include both lab-based output and non-lab-based results (e.g. radiology, physiological evaluations)</td>
</tr>
<tr>
<td>Observation EHR</td>
<td>Recording vital signs</td>
</tr>
<tr>
<td>Prescribing and drug management EHR</td>
<td>Prescribing, dispensing and inventory management purposes, as well as financial management of inventory</td>
</tr>
<tr>
<td>Documentation/clinical notes EHR</td>
<td>Documenting of clinical information inside and outside of clinical specialties</td>
</tr>
<tr>
<td>Maternal EHR</td>
<td>Management of mother and foetus during prenatal and postnatal birth stage, including charting and reporting. Once baby is born this module registers the baby with an NHS number for the rest of their life</td>
</tr>
<tr>
<td>Child health EHR</td>
<td>Vaccinations, growth charts and other algorithms and forms related to child health</td>
</tr>
<tr>
<td>Single sign on</td>
<td>Use of biometrics (e.g. facial recognition or fingerprint) or smart cards (similar to personalised bank cards) to log clinical users onto multiple disparate IT systems at automatically preventing users having to remember multiple passwords</td>
</tr>
</tbody>
</table>

**Streams and rivers: how do my patients’ data move through EHR systems?**

For clinicians, understanding data flows is key to understanding how data are shared between different EHRs. When healthcare data are entered, like a stream it can flow one way, whereas a turbulent data river can flow both ways, allowing for data exchange. Data can also differ in language and structure; therefore, if they are not standardised, the flow becomes a torrent of garbled unstructured data, useless to clinicians seeking a source of accurate, succinct and timely information (like fishing in murky waters). For communication between different EHR systems, the data need to be standardised; hence, the common data term: interoperability. Interoperability is defined as the ability of different EHR systems to share healthcare data, and is achieved through standardisation of semantics and syntax: semantic standardisation organises via language, referring to the meaning of content, whereas syntactic standardisation organises via structure, referring to the ordering of words, phrases or messages.

Semantic standardisation is traditionally achieved through clinical coding. Clinical coding is performed by GPs in GP practices, and by clinical coders in hospitals. Clinical coders read and interpret clinical text to assign relevant codes that reflects a patient’s care during a hospital visit. Correct and efficient clinical coding is essential for service evaluation, research and medical billing. Inaccuracies in coding (e.g. if a procedure that was carried out is missed) leads to the misallocation of staff, funds and resources, resulting in inefficiencies and waste. For foundation doctors, this often means the frustrating task of filling in structured proformas or checklists to help inform clinical coders before an administrative task is complete or a patient is discharged. The major coding system currently used for diseases is the International Classification of Diseases version 10 (ICD-10). Office of Population and Census survey version 4 (OPCS-4) is the coding classification used for interventions and procedures. Another newer ontological coding system is called Systematized Nomenclature of Medicine-Clinical Terms (SNOMED-CT); this covers diseases, procedures as well as many other biomedical and healthcare concepts (e.g. anatomy, occupation, etc.). Unlike other
Machine learning, artificial intelligence and data lakes: how can my patients’ data be better utilised?

Foundation doctors who use EHR certainly understand the inefficient ways in which patients’ data are utilised; however, newer innovative approaches are on the horizon, such as by using an open-source artificial intelligence (AI) framework called CogStack, which can extract data from unstandardised EHR and data silos. CogStack syphons out data, bringing them into one place: the data lake. Databases store data, data lakes represent large pooled databases and data warehouses represent organised data lakes. Data lakes contain millions of data sets, termed ‘big data’, sourced from a myriad of databases. Data that clinicians enter can be structured (ie lab results) or unstructured (ie clinical entries in the form of free text). Around 80% of big data are unstructured; this vital raw information cannot be used to a fraction of its worth. Big data require filtering: simplified into essential components (‘abstraction’), converted into a suitable format (‘transformation’) and standardised via terminology (‘harmonisation’). Filtering requires a natural language processing (NLP) system, a subset of machine learning (ML) and AI. AI uses advanced statistical approaches and computer models to mimic the organisation of the human brain. ML is a subset of AI, first introduced during the 1990s; it algorithmically derives patterns and ‘learns’ from past data without explicit programming. Deep learning is a subset of ML and uses big data to feed machines with large amounts of these raw data, developing its own representations needed for pattern recognition. For clinicians, a ML solution could provide a single user interface that is accessible across all systems simultaneously. The ML concept of NLP allows data to be ‘read’ quickly by using a combination of text-mining and terminology recognition (Fig 3). An illustration of an open-source NLP that uses big data is the medical concept annotation toolkit (MedCAT). MedCAT uses SNOMED-CT as a source of clinical terminology, but can use any standardised vocabulary.

Coding systems, SNOMED-CT has a larger, more detailed vocabulary and, therefore, it better reflects the language of clinicians. Syntactic standardisation allows for formal data structuring. For example, if a message states ‘NHS number’, the recipient knows that the next 10 digits is the NHS number; similarly, dates follow a date–month–year format. The area in healthcare that has most effectively standardised itself is radiology, via Digital Imaging and Communications in Medicine (DICOM). Given that DICOM standards were decided early, different hospitals could easily share imaging. In the rest of healthcare, the main data messaging standard is Health Level Seven (HL7); a universal packaging system and a standard of transmitting data to allow different EHR systems to communicate. The latest HL7 incorporates a flexible data structure called ‘fast healthcare interoperability resources’ (FHIR) (Fig 2). Unfortunately for foundation doctors, many EHR systems were installed without implementing adequate data syntactic standardisation because of cost, effort and inability to gain consensus; therefore, patients’ data can easily get lost or confined inside a difficult to access system. Moreover, owing to data protection legislation, healthcare data need to be retained for access for up to 25 years; therefore, data can be trapped for decades in old systems that function poorly on modern computers.

Ethical considerations of big data: how can I protect my patients’ data?

If AI is to serve as a trusted clinical aid, it must be robust, accurate and as free as possible from bias. Moreover, big data in AI

Fig 1. Frontend and backend disconnect. This represents the frontend and backend of an electronic health record. For more information is exists on the backend, than is available for use on the frontend. A scanned picture of a clinical letter may have the diagnosis and drug list and may be accessible in the frontend, but if it isn’t suitably stored in the backend (requiring specialist coding knowledge) it cannot be reused in another document like a prescription chart or a discharge summary. DHx = drug history; DOB = date of birth; MHx = medical history.
Fig 2. Data flow via packaging systems.
This represents data flow between different kinds of electronic healthcare records (EHR). Data streams flow one way and data river flow both ways allowing for data exchange. Exchange is facilitated by the packaging system HL7. Some EHR have multiple flows of data others such as Silos are completed isolated and unconnected. The names and acronyms of the different systems are given in Table 1. EHR = electronic health record; LIM = laboratory information system; PACS = radiology picture archives and communication system; PAS = patient administration system.

Fig 3. Data siphoning and NLP use.
This represents data siphoning with the open-source framework Cogstack, creating a data lake. The end point of this requires natural language processing (NLP) to clean the data and extract useful information. Example are shown of how NLP can be used. EHR = electronic health record.
represents the personal information of millions of individuals; thus, respect of privacy via consenting and prevention of data breaches via secure storage embody important safeguarding strategies.\textsuperscript{12,13} Safeguarding data by only including individuals who actively provide consent can appear fair but cause bias because certain socioeconomic classes, ethnic groups and others (eg those with a learning disability) are under-represented in consenting. Opt-out strategies, as well as ensuring openness and honesty, might be the most logical approaches to handling big data in these scenarios.\textsuperscript{13,14}

Anonymising data (all identifiers removed) or pseudonymising data (all identifiers replaced) could reduce the risk of re-identification breaches, especially in research.\textsuperscript{7} However, these strategies can lead to vital information, such as demographics, ethnicity or age, being lost, producing biases through missing data. For example, if poorer outcomes are experienced in a certain postcode but that location information is lost via anonymising, research outcomes that appear fair might in fact be biased. Bias does not imply intent; it might simply be unrepresentative of the entire patient population. A solution could be to establish a trusted research environment in which data are more openly used by trusted organisations; this was successful deployed during the coronavirus pandemic, allowing real-time data analysis in a fast-evolving situation.\textsuperscript{8,15}

\textbf{Conclusion}

As EHRs continue to grow, high-quality patient care will require foundation doctors to have a detailed understanding of digital healthcare systems. It will also become essential to teach all doctors about EHR design as well as to train them regarding how best to exploit modern data strategies to efficiently process, extract and utilise raw data obtained at the point of entry. There is a danger of creating AI that has built-in bias\textsuperscript{13,13}, however, with the hindsight of knowing how EHRs have historically worked, there is a unique opportunity to create a system built by design rather than necessity: this requires being conscious in its creation. Health inefficiencies and health inequality continue to represent huge socioeconomic burdens. Establishing EHR interoperability, instituting integrated monolithic EHRs and utilising machine learning tools represent extremely exciting future solutions to these conundrums.

\textbf{Acknowledgements}

We thank Rudy Harris, Klara De Baerdemaeker and Hani Salih for helping to sense check this piece for junior doctors and the general reader.

\textbf{References}