BMJ Open

Cohort profile for development of machine learning models to predict healthcare-related adverse events (Demeter): clinical objectives, data requirements for modelling and overview of data set for 2016–2018

Svetlana Artemova,[1,2] Ursula von Schenck,[3] Rui Fa,[4] Daniel Stoessel,[3] Hadiseh Nowparast Rostami,[3] Pierre-Ephrem Madiot,[5] Jean-Marie Januel,[2] Daniel Pagonis,[6] Caroline Landelle,[2,6] Meghann Gallouche ,[2,6] Christophe Cancé,[1,2] Frederic Olive,[6] Alexandre Moreau-Gaudry,[1,2] Sigurd Prieur,[3] Jean-Luc Bosson [1,2]

To cite: Artemova S, von Schenck U, Fa R, *et al.* Cohort profile for development of machine learning models to predict healthcare-related adverse events (Demeter): clinical objectives, data requirements for modelling and overview of data set for 2016–2018. *BMJ Open* 2023;**13**:e070929. doi:10.1136/bmjopen-2022-070929

- ► Prepublication history and additional supplemental material for this paper are available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2022-070929).

Received 09 December 2022
Accepted 27 July 2023

For numbered affiliations see end of article.

Correspondence to
Professor Jean-Luc Bosson;
Jean-Luc.Bosson@univ-grenoble-alpes.fr

ABSTRACT

Purpose In-hospital health-related adverse events (HAEs) are a major concern for hospitals worldwide. In high-income countries, approximately 1 in 10 patients experience HAEs associated with their hospital stay. Estimating the risk of an HAE at the individual patient level as accurately as possible is one of the first steps towards improving patient outcomes. Risk assessment can enable healthcare providers to target resources to patients in greatest need through adaptations in processes and procedures. Electronic health data facilitates the application of machine-learning methods for risk analysis. We aim, first to reveal correlations between HAE occurrence and patients' characteristics and/or the procedures they undergo during their hospitalisation, and second, to build models that allow the early identification of patients at an elevated risk of HAE.

Participants 143 865 adult patients hospitalised at Grenoble Alpes University Hospital (France) between 1 January 2016 and 31 December 2018.

Findings to date In this set-up phase of the project, we describe the preconditions for big data analysis using machine-learning methods. We present an overview of the retrospective de-identified multisource data for a 2-year period extracted from the hospital's Clinical Data Warehouse, along with social determinants of health data from the National Institute of Statistics and Economic Studies, to be used in machine learning (artificial intelligence) training and validation. No supplementary information or evaluation on the part of medical staff will be required by the information system for risk assessment.

Future plans We are using this data set to develop predictive models for several general HAEs including secondary intensive care admission, prolonged hospital stay, 7-day and 30-day re-hospitalisation, nosocomial bacterial infection, hospital-acquired venous thromboembolism, and in-hospital mortality.

STRENGTHS AND LIMITATIONS OF THIS STUDY

- ⇒ Data is both multisource: a broad range of clinical (eg, diagnoses, laboratory values, procedures), administrative (eg, admission mode, movements within the hospital, length of stay) and epidemiological (eg, social determinants of health like employment status, size of the family, housing conditions) data; and multitarget: we consider multiple health-related adverse events (HAEs).
- ⇒ The HAEs we consider have high impact on material and human resources.
- ⇒ Non-standardised data from the electronic medical record (EMR) has not been used in the analysis and EMR timestamps for admission, diagnoses and procedures are not precise.
- ⇒ Social determinants of health information is matched to every patient's data based on a geocoding code (the patient's home address) and is not specific for the individual.
- ⇒ Machine learning and validation needs large data sets, so only frequent HAEs can be considered.

INTRODUCTION

Informing clinical decisions is extremely important, not only for better outcomes for patients, but also for the hospital management in their efforts to reduce costs and increase efficiency. Identifying health-related adverse events (HAEs) that can be avoided and/or anticipated is a relatively straightforward step towards this improvement.

The definition of an HAE is 'harm to a patient as a result of medical care or in a health setting'.[1] HAEs are observed in 8–12% of hospitalisations within the European

Union[2–5] and in 4–24% of hospitalisations in the USA.[6 7] According to a report from the OECD (Organisation for Economic Co-operation and Development) in 2017, more than 10% of hospital expenditure is related to the treatment of HAEs that occur during patient hospitalisation.[8] These HAEs concern all surgical and medical units, including intensive care units (ICU). The main HAEs are postsurgical problems, hospital-acquired infections, venous thromboembolism, iatrogenic drug-related events, bedsores and traumatic pathologies related to care (pneumothorax, vascular access problems and foreign bodies forgotten during surgical procedures). For high-risk patients justifying particular attention at admission, one should add vital risk (death within 14 days) and secondary ICU admission.[9] The risks of excessive length of stay and unscheduled re-hospitalisations are also important elements in optimising the management of the in-hospital care pathway. Most of these elements are the target events considered in the present paper, except for drug-related adverse events and repeated surgery, because the initial analyses focus on impactful adverse events that are reliably identifiable with the available data.

We know that with the implementation of appropriate strategies, the majority of HAEs could be prevented. It has been estimated that up to 24% of all hospitalisations are affected by one or more HAE, with 23–70% potentially preventable.[7 10] Moreover, HAEs pose a significant economic burden on healthcare institutions. A study in Irish hospitals estimated that the mean length of hospital stays increased by over 6 days per HAE.[11] Nosocomial infections are an example of HAEs with a high impact on patients' well-being as well as on healthcare costs. Patient harms and HAEs consume approximately 15% of acute care expenditure in healthcare systems.[10]

Therefore, the identification of patients at risk of HAEs is essential if we wish to improve patient care and achieve cost-efficient use of overstretched human and material resources. Assessing an individual's risk of HAE as accurately as possible is the first step towards improving their safety. Such risk assessment could help healthcare providers allocate resources through adjustments in processes and procedures related to safety and quality of care. In the past, assessments of risk have been based on specific risk-by-risk scores requiring additional information to be entered by the medical team. In practice, even these validated scores are rarely used and prevention is based on the hospital's overall strategy and the experience of caregivers.

Electronic medical records (EMRs) have become an indispensable resource for clinicians when making decisions that result in improved patient outcomes. Moreover, the availability of EMR data facilitates the use of new approaches such as state-of-the-art machine-learning methods for risk prediction. In many countries Clinical Data Warehouses (CDW) collect and reuse healthcare data from EMR for various applications covering all domains of medicine.[12–20] To integrate and analyse EMR data from patients treated at our university hospital for the purposes of research, education and institutional management, the hospital has established a CDW.[21] This data can be supplemented with publicly available data from the French National Institute of Statistics and Economic Studies (INSEE) (eg, data on social determinants of health (SDOH)).

The 'development of machine learning models to predict healthcare-related adverse events – DEMETER' project (Développement de modèles de Machine Learning prédictif d'événements indésirables liés aux soins) is an international collaborative project between a university hospital and an information and analytics company aiming to improve the quality of care for patients through the use of retrospective data to make risk assessments.

Our goal is to reuse the data that are progressively recorded in the hospital's CDW throughout the treatment and follow-up of patients to optimise the healthcare pathway of all individuals admitted to the hospital in the future. There already exist many validated scores and recognised risk factors allowing clinicians to evaluate the risks of certain HAEs for a given patient, such as the risk of being admitted to the ICU if pneumonia is diagnosed or the risk of venous thromboembolic disease. We also know that these risks can be minimised by taking preventative measures. However, to correctly target patients at high risk of an HAE we first need to identify them and second, to implement appropriate prevention tailored to the nature and level of each risk. These two steps require not only the skills of individual healthcare professionals but also a common collective strategy. The characterisation of groups of high-risk patients is not only essential for each individual, but also for all patients presenting similar profiles, so that the hospital can optimise human resources and initiate preventive pharmacological or non-pharmacological interventions.

Big data analytics are particularly useful for the automated identification of HAEs from health databases.[22] Machine-learning methods have been shown to be a promising tool for risk factor identification because they enable simultaneous modelling of hundreds of variables and can reveal unsuspected correlations.[23–25] Various quality control measures have been implemented at our hospital, but with this project, for the first time, cutting-edge machine-learning tools will be used for the systematic identification of subgroups of patients at greater risk of predefined HAEs. In contrast to most previous studies, we not only consider patients with specific diseases but include the whole spectrum of hospitalised patients. Using the retrospective data, we can uncover correlations between HAE and patient characteristics and/or the procedures they undergo during their hospitalisation. This will enable us to build models to detect patients who are at an increased risk of HAE at an early stage of their hospitalisation. Standardised output and validation enable direct comparison of models to identify the optimal algorithm for the desired modelling task.

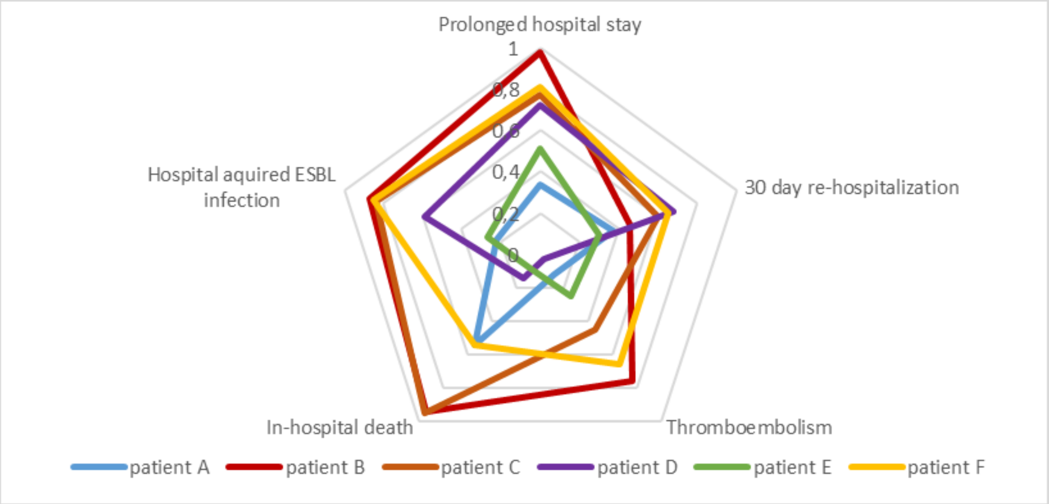

Figure 1 Example of a prediction diagram for individual patient risks estimated using machine-learning models. Five different outcomes for six patients according to the data available during the first 48 hours after admission. ESBL, extended-spectrum beta-lactamase producing *Escherichia coli* (antibiotic-resistant strains of *E. coli*).

In this project, our goal is to estimate for each individual patient the risks of a set of well-defined HAE so as to implement targeted preventive actions at the point of care. In figure 1 we show an example of the estimation of the risks of five different outcomes for six patients according to the data available during the first 48 hours after admission. This should enable us to take preventive actions and to estimate the total costs for each medical unit.

In the present paper, we describe the prerequisites for data analysis: the collection of appropriate data, the establishment of a secure analytical environment, variable selection and data preparation. We show overall results of data collected for the period 2016–2018 as an example.

METHODS

Population description

Grenoble Alpes University Hospital (France) serves a region with approximately 800 000 inhabitants, is internationally recognised as a leading centre in several medical and surgical fields and is currently ranked as the top trauma centre in the country. The agglomeration's dynamic research environment attracts many students and young highly qualified workers.

Data selection and data protection

This project uses retrospective de-identified multisource data from patients admitted to our university hospital between 1 January 2016 and 31 December 2018. For medical and legal reasons, we restrict the project database and our analysis to patients aged at least 18 years at admission and take into account only medical, surgical and obstetrical admissions (we do not include admissions in psychiatry or admissions directly to post-acute care and rehabilitation facilities).

Hospitalisations with a length of stay of less than 2 days are excluded, as well as those classified as long-stay hospice care, permanent hospitalisation, outpatient or day clinic admission (ie, admissions with a 'homogeneous group of patients' (HGP) code starting with '28' such as ambulatory chemotherapy or dialysis). Data with coding errors (HGP code starting with 90) are also excluded. In addition, patients with missing data for age or sex are excluded.

If available, we also use longitudinal patient data from previous hospitalisations in the same hospital, from up to 24 months before the current admission.

Data stored in the hospital's CDW collected from multiple sources includes patient demographics (age, sex, etc), hospitalisation details and transfers between different departments, laboratory analysis results, diagnoses, medical acts and medication prescription and administration during the hospital stay. Using the patients' residential addresses (stored in coded form in the CDW) we match this clinical data with the open-source data from the French INSEE[26] for SDOH. SDOH by submunicipal code (zones of between 1800 and 5000 inhabitants) are attributed to a subset of patients with addresses in a format usable by the geocoding platform. Patients are geocoded using the French National Address Database Geocoding Service.

To protect patients' privacy, no directly identifiable data are included in the project. We de-identify clinical data by replacing internal patient and hospitalisation codes by hash codes (an alphanumeric sequence of 40 characters generated by the SHA1 algorithm). For each study using the CDW, we generate a specific hash key to avoid any possible crossing of data and direct patient re-identification.

Selection of variables for machine-learning models

Table 1 lists all the variables selected for the project. Concerning data from our CDW, all available variables are extracted regardless of the percentage of missing values.

Table 1 List of selected variables

Individual patient information		Type
Patient demographics	Year of birth (age)	Continuous
	Sex (male, female)	Categorical
Hospitalisation	Admission date	Continuous
	Discharge date	Continuous
	Movement within the hospital	Categorical
Diagnoses	Primary, associated and related diagnoses coded using ICD-10: 1846 unique diagnoses grouped in principal and secondary diagnosis (first three characters)	Categorical
Laboratory values	Test ordered (yes, no)	Categorical
	Sampling date	Continuous
	Test result	Continuous
Medication	Prescriptions	Continuous
	Administered drugs	Continuous
Procedure codes	Classification of medical acts: 1679 unique codes (first four characters)	Categorical
ATC code	87 unique ATC codes (first three characters)	Categorical
OBS_NORM (laboratory values)	1724 unique laboratory tests: ranked as high, normal, low, tested, not tested	Categorical
Population based information by geocoding code		
Social determinants of health	Age	Continuous
	Family composition	
	Employment status	
	Living conditions	
Administrative data		
Postal Code	rural, semi-rural, urban, extraterritorial, none	Categorical
Entry mode	Eg, transfer from another establishment, after consultation with a hospital specialist, referred by an external doctor, work accident, …	Categorical
Entry type	Urgent, with confirmed appointment, spontaneous	Categorical
Exit mode	Eg, transfer to another establishment, death, return home	Categorical
Hospital clinical unit (ward) code	186 unique categories	Categorical

ATC, Anatomical Therapeutic Chemical; ICD-10, International Classification of Diseases - 10th version; OBS_NORM, measured laboratory value compared to normal laboratory test value.

The SDOH variables used for modelling were selected according to expert advice on the impact of these variables on an epidemiological study and depending on whether they were available for the general population in the latest population census (2018).

HAE examples

For risk assessment we are focussing on eight in-hospital adverse events (table 2). For each HAE a subset of the project population is selected.

Given the variable-length data provided by medical and clinical coding systems, such as International Classification of Diseases - 10th version (ICD-10), Anatomical Therapeutic Chemical (ATC) code and Classification Commune des Actes Médicaux (Common Classification of Medical Procedures in English, CCAM) codes, we implement a two-level feature engineering approach to manage the data. First, since these codes are hierarchical, we can truncate them retaining only the higher-level codes and group patients accordingly. For instance, we used only the first three characters from the ICD-10 codes, the first three characters from the ATC codes and the first four characters from the CCAM codes. This step enabled us to reduce the number of variables to a manageable amount. Second, we further reduce the number of variables by removing those with a coverage of less than 5%. By doing so, we eliminate infrequently occurring variables that may add noise to the analysis and it allows us to focus on dominant features.

Considering that missing data imputation is a complex topic and highly dependent on various factors,[27]

Table 2 Definitions of selected general hospital adverse events for risk prediction

Health adverse event	Definition	Observation window (days)*
Secondary admission to an ICU	Admission to an ICU during hospitalisation other than on the day of hospital admission or the same day as surgery.	1
Prolonged hospital stay	Hospitalisations with a length of stay (in days) above the 90th percentile for the general population are defined as prolonged.	2
30-day readmission	Re-hospitalisation within 30 days of discharge.	Entire stay
7-day readmission	Re-hospitalisation within 7 days of discharge.	Entire stay
Nosocomial bacterial infection	An associated diagnosis of significant bacterial infection at discharge, indicating infection acquired during the hospitalisation in patients admitted for other reasons.	2
Hospital-acquired venous thromboembolic disease	Hospital-acquired venous thromboembolic disease includes: pulmonary embolism and deep vein thrombosis (DVT, that can be either DVT-L=lower limb or DVT-U=upper limb).	1
In-hospital death	Adult patients who died between the 2nd and the 30th night after admission.	1
High-risk patients (in-hospital death, long length of stay and ICU secondary admission)	Adult patients who died (as defined above), were hospitalised over a long period (over 17 days) or had secondary ICU admission (as defined above).	2

*A '1-day observation window' is information collected from at least one complete day of stay, including overnight. A '2-day observational window' is information collected from at least two complete days of stay, including two nights.
ICU, intensive care unit.

including the nature of the missingness and the specific data set, we impute the missing values with zeros since the missing values in our case are not missing at random. The missing values primarily result from clinicians' decisions, reflecting the specific circumstances and clinical considerations during data collection.

Timeline used for modelling

Models that efficiently estimate the risk of 'hospital-acquired' HAEs need to capture the information available before the event occurs, at an early stage during a patient's hospitalisation. Different observation windows have been defined for the different events (table 2). To estimate an individual's risk of thromboembolism, for example, an observation window of 2 days after admission is defined. For modelling we only use the data available within this observation window.

Authorisations for the study

Our hospital's CDW was authorised by the French General Data Protection authority (GDPR) (Commission Nationale de l'Informatique et des Libertés (CNIL)) on 10 October 2019.[28] The present study was approved by the GDPR after measuring the benefit-risk ratio for the patients and examining the security measures put in place. Eligible patients were not individually informed about the study because of their very large number. General information about the study is available on Grenoble Alpes University Hospital's website.[29] Patients have the right to refuse that their individual de-identified data are included in the CDW.

Patient and public involvement statement

To date there is no patient or general public involvement in the study; however, details of the CDW and the study are available on the hospital's website. We would be happy to consider suggestions from patients, patient associations or the general public about future studies that could be done using the methods we have developed and the database (updated if necessary).

Analytic environment and its security

To ensure patient data confidentiality, de-identified and tokenised patient data selected for the study was stored on a secure server, with access through a virtual desktop and permanent desktop monitoring for accredited users. The server uses two AMD APYC 7742 64 core processors with 1 TB of memory and Dell express flash NVMe P4610 3.2 TB SFF storage. In addition, the server includes an NVIDIA V100 Tensor Core GPU. Windows Server 2019 Datacenter V.1809 was installed on the server and Anaconda V.4.10.1, Jupyter Notebook V.6.3.0 and Python V.3.8.8 are used for data processing. We also installed and configured Microsoft SQL Server 2019 and NVIDIA CUDA Deep Neural Network library. In the online supplemental material eFigure 1 illustrates the security concept and remote data access.

Project cost

The total cost of the project set-up was:

- Acquisition of the server: €30 000.
- Hospital programmers (2 years part-time, ie, 0.6 person-years).

- Data scientists and medical expert (2 years part-time, 0.3 person-years).

Collaboration

The DEMETER project is currently a public–private collaboration. The present paper gives an overview of the initial data set to be used for machine learning and validation. Analyses concerning several general HAE (eg, mortality, length of hospital stay) based on the current data set are nearing completion.

As a next step we plan to extend the project and include three more university hospitals within France and to re-validate the models for the extended data set.

CHARACTERISTICS OF THE DATA SET

Composition of database (2016–2018)

Initially, data on 545 628 hospital admissions during the period between 1 January 2016 and 31 December 2018, concerning 237 657 individuals, were available in our CDW (see online supplemental material eTable 1). After applying all the filters (online supplemental eTable 1) data on 123 729 hospital admissions (79 117 individuals) met the requirements to be included in the data set for machine learning. Many patients were admitted several times during the period used for the data set (online supplemental eTable 1). An overview of the characteristics of the patients and admissions retained in the data set is presented in table 3.

Table 3 Characteristics of patient admissions in the final data set for the period 2016–2018

Characteristics of admissions retained in data set	Admissions n (±SD)	Admissions %
Age (mean±SD) (years)	61±20.51	Na
Sex, female	62 156	50.23
Length of stay (mean±SD) (days)	7.88±9.49	Na
Emergency admission (first hospital code: ICU)	1076	0.87
Urgent admission	7834	6.33
Urban postal code	66 571	13.66
Rural postal code	16 909	13.66
Admissions with at least one diagnosis	123 729	100
Admissions with at least one medication	120 380	97.29
Admissions with at least one procedural code	113 802	91.98
Admissions with at least one clinical unit code	123 729	100
Number of distinct clinical unit codes (mean±SD)	1.72±0.88	Na
Number of distinct diagnoses (mean±SD)	9.49±7.20	Na
Number of distinct procedure codes (mean±SD)	5.18±4.56	Na
Number of distinct ATC codes (mean±SD)	6.63±3.81	Na

ATC, Anatomical Therapeutic Chemical; ICU, intensive care unit; Na, not applicable.

Geocoding

Geocoding resulted in the positioning of 93.3% patients out of 142 364, for whom a geocoding code for where they lived could be attributed (online supplemental eFigure 2), and 77.3% of patients (out of 142 364) had a geocoding assessment score free of positioning errors. We used a Threshold Assessment Protocol to evaluate the credibility of the patients' addresses. Currently the address credibility threshold score is 51.7. After 1109 manual address improvements 79.6% out of the 23 301 patients from the agglomeration were geocoded with a score above the chosen threshold and associated with a geocoding-code while 816 were not geocoded. In patients from other places, mostly rural areas, the geocoding code is determined according to the name of the village.

SDOH by geocode

Online supplemental eTable 2 shows SDOH for the geocoded areas in which the general study population lived.

Laboratory tests

We categorise hospital analytical laboratory values as high, normal and low, and as 'tested' or 'not tested', to account for the physicians' decision to ask for a laboratory test. Laboratory results were available for over 80% of admissions in the period studied (table 4).

DISCUSSION

Project originality

A growing number of studies evaluate the use of artificial intelligence for the improvement of the quality of care and the prevention of HAEs. In contrast to our study, most of them focus on very specific medical conditions, for example, an automatic analysis of ECGs to detect asymptomatic left ventricular dysfunction after 65 years of age, which appears to be cost-effective[30] or in-hospital mortality.[31] The methodological originality of our project is the fact that it is both multisource because we use a broad range of clinical, administrative and epidemiological data and multitarget as we consider multiple HAEs.

We have implemented a process to assess the risks of several common HAEs in a broad population representative of those served by a large university hospital, especially in terms of patient complexity and length of stay. The HAEs we considered in this first step were selected because of their impact (being burdensome in terms of material and human resources), frequency, ease to define

Table 4 Analytical laboratory tests

Laboratory tests (ordered at least once)	Admissions n (±SD)	Admissions %
At least one laboratory test	115 677	93.49
Blood count (haematocrit)	104 900	90.60
Electrolytes (sodium)	94 587	81.69
Blood glucose	52 492	45.33
Blood creatinine	94 058	81.23
Liver enzymes (alanine aminotransferase)	65 007	56.14
Prothrombin time	74 998	64.77
C-reactive protein	75 879	65.53
Bicarbonate	88 248	76.21

and either preventable or important for care planning (eg, hospital bed management). Although of interest, very common drug-related adverse events were not considered in this first step. Drug-related adverse events are too difficult to identify automatically outside their individual analysis by a pharmacovigilance expert.

The project is only possible thanks to the hospital's CDW[21 32] that compiles clinical as well as administrative data from various sources. We use all available data from the first or the first 2 days of a patient's hospitalisation to predict their risk for HAEs. There is a trade-off between the timeliness of the prediction and the data available for modelling.

Since the massive deployment of prescription software, drug prescriptions are a major source of information that is rapidly available in the hours following a patient's admission to hospital. Data based on prescriptions is valuable and provides information on a patient's comorbidities (eg, insulin and diabetes) and condition (type, quantity and administration routes of drugs).[33]

In contrast to clinical trials, in the real-world setting not every patient has a test result for every laboratory test or procedure prescribed. Most studies either regrade tests that have not been performed as missing values or impute data. We use the information that a specific test or procedure is ordered, or is not ordered, as a separate variable, because in a real-world setting a test or procedure is only ordered if the physician considers it will have some impact.

The fact that SDOH has an impact on clinical outcomes is widely accepted. Nevertheless, most predictive models are based solely on clinical data. Adding administrative information (eg, the admission mode) and SDOH (eg, unemployment rates in the patient's district of residence) to our database is a huge advantage of our study.

Assessment of the various risks of adverse events that may occur during hospitalisation give added value to data collected in routine everyday practice and is invaluable for both the medical teams in direct contact with patients and also for overall hospital management. At the individual level, each team has greater or lesser expertise in assessing and dealing with certain risks. The presentation of all possible risks for patients could optimise risk management by strengthening the weak points of the care teams. At the level of the establishment as a whole, the challenge is to optimise the necessarily limited means of preventing iatrogenic risks by directing these means towards patients and departments at high risk. Mapped across the entire hospital and reassessed daily, would also allow the better orientation of new patients (to avoid saturating a department with a high iatrogenic risk) and better matching of staffing needs.

Although setting up this project was costly it is an investment as it can be used for further research. The server can be used for other machine or deep learning projects, and the database is a rich source for further analysis of this general population. The knowledge about the data and its quality that we have acquired during this project can be used for future research.

Project challenges and limitations

Our study was initiated in late 2019. Acquiring regulatory approval was delayed due to the COVID-19 pandemic. It took some time to establish processes compliant with legal requirements and to carefully review the databases (data source: when and how was the information entered; potential biases; understanding variations in data completeness, etc).

We use real-world retrospective de-identified patient data and do not include any of the textual data available in the CDW. Including information from clinical notes would require language processing and additional de-identification. Some information available to the physician, for example, results from bedside fingertip blood glucose tests are not available in the CDW. Diagnoses coded for invoicing purposes do not need to be complete, for example, not all comorbidities need to be coded to invoice a specific surgical procedure. Thus, we have a bias related to the activity-based pricing model of hospital funding. Furthermore, diagnosis codes for invoicing purposes do not need precise timestamps. Therefore, sometimes indirect information, for example, from laboratory tests must be used to verify that an event is 'hospital-acquired' and was not present at admission. The observation window for data collection was defined in 'days' rather than hours because of delays in entering information into the hospital information system.

Nationally, there is no computerised medical record shared between hospitals, therefore historical data on our patients is incomplete.

In order to train reliable models, we need at least a few hundred cases. Machine learning models are thirsty for large data, not only because large amounts of independent data may bring better statistical estimates, but also because the rigorous training and evaluation process restricts the whole database from being used in the training process so as to leave a portion aside for evaluation to generalise the model. Therefore, only frequent events are included in the study.

SDOH information based on a geocoding code is not specific for an individual patient but for the population of a patient's home district. In addition, the population of our region is not very diverse, the variance within SDOH is much lower than for our country as a whole or across Europe.

For modelling we only use information available at the beginning of a patient's hospitalisation, because the assessment of risk needs to be available as early as possible to efficiently adapt the patient's care pathway. Prediction using pre-hospitalisation data would be a much more ambitious project and certainly of immense interest. It would depend not only on a patient's previous hospitalisations in our hospital but also on the patients' electronic health record with data from other establishments, general practitioners, pharmacies, medical imaging centres, private analytical laboratories, physiotherapists, etc.

Ongoing and future studies

We are using the 2016–2018 data set to train risk assessment models for eight different HAEs: secondary admission to an ICU, prolonged hospital stay, 30-day readmission, 7-day readmission, nosocomial bacterial infection, hospital-acquired venous thromboembolic disease, in-hospital death, very high-risk patient (in-hospital death, long length of stay and secondary admission to ICU).

For in-hospital mortality three out of eight different machine learning algorithms have demonstrated high discrimination with a balanced accuracy. Models are trained with 283 variables including age, sex, sociodeterminants of health, laboratory tests, procedures, medications, ward and home address. Results are evaluated using various performance metrics. We included 3542 admissions with in-hospital mortality and 120 187 admissions as controls.

In another study, applying machine learning algorithms to the 2016–2018 data set has enabled us to propose a list of the factors that are foremost in the prediction of long length of hospital stay.

In a next step we will validate all predictive models developed on more recent data from Grenoble Alpes University Hospital before we test them for individual risk prediction at the point of care.

We then plan to extend the analysis including more hospitals to evaluate whether the models can be applied for risk prediction in a more diverse population.

Conclusion

This project is a successful public–private clinical research cooperation using big multisource data. Here we have described the data available, the HAE's of interest in most large hospitals and how the data is managed in compliance with regulatory requirements. Data processing, standardisation and quality assurance of data collected only for clinical or administrative reasons is one of the greatest challenges for analytical projects exploiting hospital data. It is well known that access to real-world health data is not straightforward with questionable data quality.[34] Due diligence is required to avoid being swamped by the data available in a CDW. The CDW database and its quality continuously evolve, as new data sources are added, for example, with the more widespread availability of digital health data. The project provides the foundations for many further research projects on the quality of care.

Author affiliations
[1]Public Health Department, INSERM CIC1406, CHU Grenoble Alpes, Grenoble, France
[2]TIMC, CNRS UMR5525, Université Grenoble Alpes, Grenoble, France
[3]Life Science Analytics, Elsevier BV, Berlin, Germany
[4]Elsevier Health Analytics, London, UK
[5]Digital Services Management, CHU Grenoble Alpes, Grenoble, France
[6]Public Health Department, CHU Grenoble Alpes, Grenoble, France

Acknowledgements We thank Grenoble Alpes University Hospital management and Elsevier management (Jocelyn Dutil, Anne Kittler, Georges-Henri Lion, Olaf Lodbrok, Alice Mogenet, Ulrich Schieborr) for their support. Celine Ginet, Brigitte Cohard and Jean François Blatier for acquisition of data; and Alison Foote (an independent medical writer based in Grenoble, France) for critical reading and very substantial editing of the manuscript.

Contributors Guarantor: J-LB is responsible for the overall content, accepts full responsibility for the work and/or the conduct of the study, had access to the data and oversaw the decision to publish. SA: Study design and planning, acquisition of data, drafting the article (literature review, discussion of the impact of study and its result), approval of the version to be submitted and agreement to be accountable for all aspects of the work. UvS: Study design and planning, drafting the article (introduction to the topic, literature review, discussion of the impact of study and its result), approval of the version to be submitted and agreement to be accountable for all aspects of the work. RF: Study design, data analysis, drafting the article (analytical methodology, visualisation), approval of the version to be submitted and agreement to be accountable for all aspects of the work. DS: Study design, data analysis, drafting the article (secure data handling, categorisation), approval of the version to be submitted and agreement to be accountable for all aspects of the work. HNR: Study design, data analysis, literature review, critical revision of manuscript, approval of the version to be submitted and agreement to be accountable for all aspects of the work. P-EM: Study design and planning, acquisition of data, data management, critical revision of manuscript, approval of the version to be submitted and agreement to be accountable for all aspects of the work. J-MJ: Study design, definitions of adverse events, drafting the article (introduction to the topic, discussion of the impact of study and its result), approval of the version to be submitted and agreement to be accountable for all aspects of the work. DP: Study design, coordination of study, supervision of data extraction and filtering, assistance with data security aspects, critical revision of manuscript, approval of the version to be submitted and agreement to be accountable for all aspects of the work. CL: Study design, acquisition of laboratory analysis data, interpretation of results, drafting the article (discussion of the impact of study and its result), approval of the version to be published and agreement to be accountable for all aspects of the work. MG: Study design, acquisition of data, definitions of adverse events, literature search, critical revision of manuscript, approval of the version to be submitted and agreement to be accountable for all aspects of the work. CC: Study design, construction of data set, data analysis (geolocalisation data), critical revision of manuscript, approval of the version to be submitted and agreement to be accountable for all aspects of the work. FO: Study design, assistance with the interpretation and use of CIM 10 data and classification of procedures, critical revision of manuscript, approval of the version to be submitted and agreement to be accountable for all aspects of the work. AM-G: Conception and design of the clinical data warehouse, assistance with ethical and regulatory aspects, construction of data set, discussion of the impact of study and its result, critical revision of manuscript, approval of the version to be submitted and agreement to be accountable for all aspects of the work. SP: Study design,

supervision of the project, discussion of the impact of study and its result, critical revision of manuscript, approval of the version to be submitted and agreement to be accountable for all aspects of the work. J-LB: Initiator of the original idea for the project, study design, acquisition of data, supervision of analyses and assistance in the interpretation of results, coordination of project, drafting the article (introduction to the topic, literature review and discussion of the impact of study and its result), approval of the version to be published and agreement to be accountable for all aspects of the work.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Map disclaimer The inclusion of any map (including the depiction of any boundaries therein), or of any geographical or locational reference, does not imply the expression of any opinion whatsoever on the part of BMJ concerning the legal status of any country, territory, jurisdiction or area or of its authorities. Any such expression remains solely that of the relevant source and is not endorsed by BMJ. Maps are provided without any warranty of any kind, either express or implied.

Competing interests None declared.

Patient and public involvement Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

Patient consent for publication Not applicable.

Ethics approval The study was approved by the French national ethics committee (Expert committee for research, studies and evaluations in the field of health/ Comité d'expertise pour les recherches, les études et les évaluations dans le domaine de la santé (CEREES) - National Institute of Health Data (INDS) n° 926994) on 2 April 2020.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement No data are available. The data in this cohort data set are not currently publicly available as they contain the complete individual medical files of all patients included in the data set, albeit de-identified, so for ethical and data privacy reasons they cannot be exported to others. Nevertheless, procedures regarding the conditions and accessing of selected de-identified data for academic use by third parties (with a signed data access agreement) are currently in discussion. Requests should be made to Grenoble-Alpes University Hospital (contact Professor Jean-Luc Bosson: jlbosson@chu-grenoble.fr).

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

ORCID iDs

Meghann Gallouche http://orcid.org/0000-0002-3135-1423
Jean-Luc Bosson http://orcid.org/0000-0003-0967-6026

REFERENCES

1 Levinson DR. Report no.OEI-06-09-00090. *Adverse events in hospitals: national incidence among Medicare beneficiaries*. USA: Department of Health and Human Services Office of the Inspector General, 2010.

2 Soop M, Fryksmark U, Köster M, *et al*. The incidence of adverse events in Swedish hospitals: a retrospective medical record review study. *Int J Qual Health Care* 2009;21:285–91.

3 Aranaz-Andrés JM, Aibar-Remón C, Vitaller-Murillo J, *et al*. Incidence of adverse events related to health care in Spain: results of the Spanish national study of adverse events. *J Epidemiol Community Health* 2008;62:1022–9.

4 de Vries EN, Ramrattan MA, Smorenburg SM, *et al*. The incidence and nature of in-hospital adverse events: a systematic review. *Qual Saf Health Care* 2008;17:216–23.

5 Magdelijns FJH, Stassen PM, Stehouwer CDA, *et al*. Direct health care costs of hospital admissions due to adverse events in the Netherlands. *Eur J Public Health* 2014;24:1028–33.

6 Brennan TA, Leape LL, Laird NM, *et al*. Incidence of adverse events and negligence in hospitalized patients: results of the Harvard Medical Practice Study I. *N Engl J Med* 1991;324:370–6.

7 Bates DW, Levine DM, Salmasian H, *et al*. The safety of inpatient health care. *N Engl J Med* 2023;388:142–53.

8 OECD. *Tackling Wasteful Spending in Health Care*. Paris: OECD Publishing,

9 Schwendimann R, Blatter C, Dhaini S, *et al*. The occurrence, types, consequences and preventability of in-hospital adverse events - a scoping review. *BMC Health Serv Res* 2018;18:521.

10 Slawomirski L, Auraaen A, Klazinga N. The Economics of patient safety: strengthening a value-based approach to reducing patient harm at national level. OECD; 2017. Available: https://www.oecd.org/health/health-systems/The-economics-of-patient-safety-March-2017.pdf [Accessed Nov 2022].

11 Rafter N, Hickey A, Conroy RM, *et al*. The Irish National Adverse Events Study (INAES): the frequency and nature of adverse events in Irish hospitals-a retrospective record review study. *BMJ Qual Saf* 2017;26:111–9.

12 Karami M, Rahimi A, Shahmirzadi AH. Clinical data warehouse: an effective tool to create intelligence in disease management. *Health Care Manag (Frederick)* 2017;36:380–4.

13 Goers R, Coman Schmid D, Jäggi VF, *et al*. Swisspk$^{\text{Cdw}}$- a clinical data warehouse for the optimization of pediatric dosing regimens. *CPT Pharmacometrics Syst Pharmacol* 2021;10:1578–87.

14 Fleuren LM, Dam TA, Tonutti M, *et al*. The Dutch Data Warehouse, a multicenter and full-admission electronic health records database for critically ill COVID-19 patients. *Crit Care* 2021;25:304.

15 Rinner C, Gezgin D, Wendl C, *et al*. A clinical data warehouse based on OMOP and I2B2 for Austrian health claims data. *Stud Health Technol Inform* 2018;248:94–9.

16 Loput CM, Saltsman CL, Rahm RC, *et al*. Evaluation of medication administration timing variance using information from a large health system's clinical data warehouse. *Am J Health Syst Pharm* 2022;79:S1–7.

17 Agapito G, Zucco C, Cannataro M. COVID-WAREHOUSE: a data warehouse of Italian COVID-19, pollution, and climate data. *Int J Environ Res Public Health* 2020;17.

18 Lamer A, Demay A, Marcilly R. Data Reuse through anesthesia data warehouse: searching for new use contexts. *Stud Health Technol Inform* 2018;255:102–6.

19 Kortüm KU, Müller M, Kern C, *et al*. Using electronic health records to build an ophthalmologic data warehouse and visualize patients' data. *Am J Ophthalmol* 2017;178:84–93.

20 Lelong R, Soualmia LF, Grosjean J, *et al*. Building a semantic health data warehouse in the context of clinical trials: development and usability study. *JMIR Med Inform* 2019;7.

21 Artemova S, Madiot P-E, Caporossi A, *et al*. PREDIMED: clinical data warehouse of Grenoble Alpes University Hospital. *Stud Health Technol Inform* 2019;264:1421–2.

22 Musy SN, Ausserhofer D, Schwendimann R, *et al*. Trigger tool-based automated adverse event detection in electronic health records: systematic review. *J Med Internet Res* 2018;20.

23 Yang Z, Huang Y, Jiang Y, *et al*. Clinical assistant diagnosis for electronic medical record based on Convolutional neural network. *Sci Rep* 2018;8.

24 Yoo KD, Noh J, Lee H, *et al*. A machine learning approach using survival Statistics to predict graft survival in kidney transplant recipients: a multicenter cohort study. *Sci Rep* 2017;7:8904.

25 Fagerström J, Bång M, Wilhelms D, *et al*. Lisep LSTM: a machine learning algorithm for early detection of septic shock. *Sci Rep* 2019;9:15132.

26 2018 census: results for zones, databases and detailed files [L'Institut national de la statistique et des études économiques]. Available: https://www.insee.fr/fr/information/5369871 [Accessed 4 Dec 2022].

27 Bodenhofer U, Haslinger-Eisterer B, Minichmayer A, *et al*. Machine learning-based risk profile classification of patients undergoing elective heart valve surgery. *Eur J Cardiothorac Surg* 2021;60:1378–85.

28 Délibération N° 2019-124 Du 10 Octobre 2019 [Commission Nationale de l'Informatique et des Libertés]. Available: https://www.legifrance.gouv.fr/cnil/id/CNILTEXT000039292717/ [Accessed 4 Dec 2022].

29 [Clinical data warehouse] Entrepôt de Données de Santé (EDS) [clinical data warehouse] projects: Demeter [CHU-Grenoble-Alpes]. n.d. Available: https://www.chu-grenoble.fr/patients-et-accompagnants/la-recherche-au-chuga/entrepot-de-donnees-de-sante-eds

30 Tseng AS, Thao V, Borah BJ, *et al*. Cost effectiveness of an electrocardiographic deep learning algorithm to detect asymptomatic left ventricular dysfunction. *Mayo Clin Proc* 2021;96:1835–44.

31 Seki T, Kawazoe Y, Ohe K. Machine learning-based prediction of in-hospital mortality using admission laboratory data: a retrospective, single-site study using electronic health record data. *PLoS One* 2021;16.

32 Otokiti A. Using Informatics to improve healthcare quality. *Int J Health Care Qual Assur* 2019;32:425–30.

33 Lepelley M, Genty C, Lecoanet A, *et al*. Electronic medication regimen complexity index at admission and complications during hospitalization in medical wards: a tool to improve quality of care? *Int J Qual Health Care* 2018;30:32–8.

34 Kim HS, Kim DJ, Yoon KH. Medical big data is not yet available: why we need realism rather than exaggeration. *Endocrinol Metab (Seoul)* 2019;34:349–54.