Bayesian networks to identify potential high-risk multimorbidity and intervention clusters in inpatients: an explorative data mining study

Roth Jan A.abc*, Sakoparnig Thomascde*, Gerber Martinf, Hug Balthaslrcq, the PATREC Study Group*

* Basel Institute for Clinical Epidemiology and Biostatistics, University Hospital Basel, Switzerland
b Division of Infectious Diseases and Hospital Epidemiology, University Hospital Basel, Switzerland
c University of Basel, Switzerland
d Focal Area of Computational and Systems Biology, Biozentrum, University of Basel, Switzerland
e SIB Swiss Institute of Bioinformatics, Lausanne, Switzerland
f Department of Finance, University Hospital Basel, Switzerland
g Department of Internal Medicine, Kantonsspital Luzern, Switzerland
h PATREC Study Group (in alphabetical order): Christian Abshagen (University Hospital Basel, Switzerland); Geoffrey Fucile (University of Basel, Switzerland); Martin Gerber (University Hospital Basel, Switzerland); Balthasar L. Hug (Kantonsspital Luzern, Switzerland); Bernd Jaegle (University Hospital Basel, Switzerland); Eleonore Kuenzel-Pawlik (University Hospital Basel, Switzerland); Simon Neubauer (University Hospital Basel, Switzerland); Rakesh Padiyath (University Hospital Basel, Switzerland); Jan A. Roth (University Hospital Basel, Switzerland); Thomas Sakoparnig (University of Basel, Switzerland); Thierry Sengstag (University of Basel, Switzerland); Damian Splyra (University Hospital Basel, Switzerland); Andreas F. Widmer (University Hospital Basel, Switzerland)

Aims

AIMS OF THE STUDY: Based on large sets of routine hospital data from inpatient cases, we aimed to explore multimorbidity and intervention clusters showing high risks for in-hospital mortality and unplanned readmissions using data-driven analytical methods.

METHODS: We performed an explorative, historical cohort study of consecutive inpatient cases at a tertiary care centre with an integrated platform for routine healthcare data in Switzerland. From January 2012 through to December 2017, all inpatients aged ≥18 years at hospital admission were eligible for study inclusion. We predefined all-cause in-hospital death and unplanned hospital readmission as co-primary outcomes. In a first step, we explored and visualised multimorbidity and intervention clusters using mutual information analysis. In a subsequent step, we trained multi-layer Bayesian networks to identify clusters associated with in-hospital death and/or unplanned hospital readmission.

RESULTS: Among 190,837 inpatient cases, 7994 unique interventions and 6639 interventions were routinely recorded during the six-year study period. Based on the mutual information analysis, we identified 32 multimorbidity clusters and 24 intervention clusters – of which several were directly related to in-hospital mortality and/or unplanned readmission in the subsequent Bayesian network analysis.

CONCLUSIONS: Bayesian network analysis may be used as a tool to mine large healthcare databases in order to explore intervention targets for quality improvement programmes. However, the resulting associations should be substantiated in consecutive investigations using specific causal models. (Trial registration no EKNZ 2016-02128.)

Keywords: Bayesian network, cluster, digital epidemiology, electronic medical record, high-risk

Introduction

A main goal of healthcare providers should be to deliver high-quality patient care and to continuously improve preventive and treatment services where needed [1, 2]. At present, many quality improvement and patient safety programmes focus on a few specific patient groups, health outcomes or high-risk interventions (e.g., prevention of postoperative infections, pressure ulcers), without any overarching, hospital-wide systematic prevention strategy. The increasing digitisation and integration of healthcare data may provide ideal conditions for the use of routine healthcare databases to identify high-risk patient and intervention groups and to define data-driven prevention targets [3, 4]. Clusters or groups of patients may be defined based on any relevant characteristic, such as specific disease combinations (multimorbidity clusters) or networks of therapeutic interventions (intervention clusters).

Using the large healthcare database of a tertiary care institution and based on previous work on multimorbidity clustering [5], we aimed to explore high-risk multimorbidity and intervention clusters by using data mining methods, namely mutual information and Bayesian network analysis, to assess the interrelationship between inpatient characteristics, diagnoses and interventions and health outcomes (unplanned readmissions and in-hospital mortality). We hypothesised that Bayesian networks may be useful...
tools with which to explore high-risk inpatient groups based on routine data from electronic medical records.

Materials and methods

Study design and patient selection

We performed an explorative, historical cohort study at the University Hospital Basel, an 850-bed tertiary care hospital in Switzerland with more than 1,000,000 ambulatory patient contacts and over 37,000 inpatient cases per year. Treatment services comprise all main medical and surgical disciplines, including hematopoietic stem cell and kidney transplantations.

From January 2012 through to December 2017, all consecutive inpatients aged ≥18 years at hospital admission were eligible for study inclusion. We excluded patients who declined to participate in observational studies that make secondary use of their routine health data (i.e. general research consent). We defined inpatients as patients who were hospitalised for at least 24 hours.

The Ethics Committee of Northwestern and Central Switzerland approved this study with a waiver of informed consent (project number 2016-02128). We reported our study according to the ‘REporting of studies Conducted using Observational Routinely-collected health Data’ (RECORD) recommendations [6].

Data extraction and study definitions

A data manager extracted the relevant study data of the included patients on an inpatient case level (for each hospitalisation period, lasting from hospital admission to discharge) from an in-house, in-memory data platform (SAP® HANA; SAP AG, Walldorf, Germany), as described previously [5, 7]. No filtering was required, as the relevant variables were available for all eligible inpatient cases. No linkage of the data with other databases was performed.

The investigators had full access to the complete database population.

For each inpatient case, diagnoses and interventions (diagnostic, medical and surgical interventions) were routinely coded by medical coders for billing and controlling purposes according to the International Classification of Diseases, 10th revision (ICD-10; German modification) and the Swiss classification of medical interventions (“Schweizerische Operationsklassifikation”) respectively.

The active discharge diagnoses of each inpatient case, which were either present on hospital admission (e.g., previously diagnosed chronic diseases) or newly diagnosed during the hospital stay, were coded [5].

Based on their clinical relevance, we defined all-cause in-hospital death and unplanned hospital readmission as a priori co-primary outcome measures. Unplanned readmissions (to the index hospital) due to the same diagnosis as during the previous hospitalisation (e.g., surgical site infection after heart surgery) are routinely coded for inpatients in Switzerland if they occur within 18 days after discharge [8].

Statistical analysis

We conducted the computations at the High Performance Computing Centre of the University of Basel. We performed all statistical analyses using the programming language Python, version 3.6.4 (Python Software Foundation, Wilmington, DE). Analysis codes can be requested from the corresponding author.

Mutual information cluster analysis

In order to quantify the occurrence relationship between pairs of diagnoses and interventions, and building upon our previous work [5], we calculated pairwise mutual information scores for all pairs of diagnoses and all pairs of interventions (diagnostic, medical and surgical interventions combined). Mutual information is an information theory concept that quantifies the general interdependence of two random variables [9]: it quantifies how much information about one random variable can be obtained through the other.

We used these mutual information scores to graph separate multimorbidity/diagnosis and intervention clusters, with diagnoses and interventions as vertices and mutual information scores as edges. Based on clinical opinion and to avoid over-clustering, we considered only mutual information scores >0.01 and we chose to cluster diagnoses and interventions separately. Mutual information cut-offs ≤0.01 led to a multitude of clinically implausible diagnosis and intervention pairs/patterns.

We identified separate clusters by looking for connected components in the mutual information graphs. We considered any inpatient to be part of a given cluster if they had at least one diagnosis or intervention from that cluster. Therefore, inpatients may have had more than one cluster association.

Learning Bayesian networks

Bayesian networks are directed acyclic graphs in which the vertices represent variables and the edges/arrows represent conditional dependencies. We included the following inpatient variables in our analysis: discretised age category at hospital admission (years; <40, 40–59, 60–79, ≥80), emergency admission status (yes/no), sex (male/female), and the discretised number of secondary ICD-10 diagnoses/comorbidities (<3, 3–9, ≥10). The respective strata were defined a priori based on consensus within the study group. For each multimorbidity/diagnosis and intervention cluster, we learned a separate Bayesian network by adding the respective cluster and outcome status.

For probabilistic modelling, we used the Python package Pomegranate to learn the Bayesian networks [10]. Based on clinical opinion and for better interpretability (with the level of complexity limited due to the explorative nature of these analyses), we added a few a priori structure constraints: the inpatient variables age, emergency admission status, sex and number of secondary diagnoses/comorbidities were restricted to layer one. Furthermore, the cluster association/membership and the outcome variable were restricted to layers two and three respectively. Only connections from layer one to layer two and/or three and from layer two to layer three were allowed. The rationale behind these restrictions was that we were interested in whether age, emergency admission status, sex and number of secondary diagnosis have an effect on cluster membership and, furthermore, whether these variables combined have an effect on the outcome. Any effects of the outcome on cluster membership were not of interest, since there cannot
be a causal effect in this direction. We used the “exact” algorithm of the Bayesian network class of the Pomegranate package, i.e. an exhaustive search over all possible structures, to learn the Bayesian networks. This procedure optimises the minimum description length (MDL) score. The respective conditional probability table is provided in the supplementary table in appendix 4.

Results

During the six-year study period there were 198,972 inpatient cases overall. After excluding 8135 cases who declined to give their general research consent, we included the remaining 190,837 cases (114,651 unique inpatients) in the final analysis. During the study period, 7994 unique diagnoses and 6639 interventions were recorded. The main characteristics of the study population are presented in table 1. The median age of the patient cases was 63 years (interquartile range 45–76), with a median length of hospital stay of 4 days (interquartile range 2–8). 3834 out of the 190,837 patient cases (2.0%) died during the hospital stay and 15,643 out of the 190,837 patient cases (8.2%) had an unplanned hospital readmission.

We identified 32 diagnosis/multimorbidity clusters (overview in fig. 1A) and 24 intervention clusters (overview in fig. 1B) in the mutual information analysis. These clusters covered a wide spectrum of diagnoses as well as medical and surgical interventions. The respective clusters are described in detail in the supplementary text (appendix 1) and the figures in appendices 2 and 3. Several of these clusters were found to be directly related to in-hospital mortality and/or unplanned readmissions in the subsequent Bayesian network analysis. Furthermore, various multimorbidity and intervention clusters were only indirectly related to in-hospital mortality and unplanned readmissions via patient and admission characteristics (age, sex, emergency admission status and number of comorbidities). The Bayesian networks for all multimorbidity and intervention clusters are presented in the supplementary materials (appendices 2 and 3). For all-cause in-hospital death and unplanned readmission, we chose, for illustrative purposes and for better interpretability, to describe eight Bayesian networks, which are depicted in figures 1C to 1F.

The diagnosis cluster presented in figure 1C (cluster 8) consists of three diagnoses, describing a patient group with incidents caused by medical devices and products (ICD-10 Y82.8), with infections or inflammatory reactions through prostheses/implants in the urinary tract (ICD-10 T83.5), or with mechanical complications of a joint endoprosthesis (ICD-10 T84.0). Overall, 6,107 out of 190,837 cases (3.2%) had at least one of these three diagnoses during their hospital stay. In the corresponding Bayesian network, a direct relationship between cluster 8 and in-hospital death was observed, and the cases’ age, sex, emergency admission status and number of comorbidities were directly linked with the probability of belonging to cluster 8. There was, however, no direct link between the cluster and unplanned hospital readmissions. In cluster 8, 160 out of 6107 (2.6%) patients died during the hospital stay and 1140 out of 6107 patients (18.7%) had an unplanned hospital readmission. The patients’ underlying characteristics (age, sex, emergency admission status and number of comorbidities), but not the multimorbidity cluster association, seem to explain the high readmission proportion.

Figure 1D depicts a large neurological and diagnostic cluster (cluster 4) consisting of 19 interrelated interventions and diagnostic procedures. These ranged from computed tomography of the abdomen (87.41.99 and 33.24.13) and thorax (87.41.99) to different tracheobronchoscopic interventions (33.24.13, 33.24.11 and 33.24.14) and neurological diagnostics and treatments. Overall, 33,942 out of 190,837 cases (17.8%) had at least one of these diagnoses during their hospital stay. In the respective Bayesian network, a direct relationship between cluster 4 and in-hospital death was observed. As in cluster 8, there was no direct link with unplanned hospital readmissions. In cluster

Table 1: Characteristics of the study population (n = 190,837 inpatient cases).

<table>
<thead>
<tr>
<th>Variable</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of unique inpatients</td>
<td>114,651</td>
</tr>
<tr>
<td>No. of cases per inpatient during the study period, mean</td>
<td>1.66</td>
</tr>
<tr>
<td>Case characteristics</td>
<td></td>
</tr>
<tr>
<td>Age in years at hospital admission, median (IQR)</td>
<td>63 (45–76)</td>
</tr>
<tr>
<td>Female sex, n (%)</td>
<td>97,521 (51.1)</td>
</tr>
<tr>
<td>Length of hospital stay, median (IQR)</td>
<td>4 (2–8)</td>
</tr>
<tr>
<td>Emergency admission, n (%)</td>
<td>90,969 (47.7)</td>
</tr>
<tr>
<td>No. of secondary diagnoses/comorbidities⁴, median (IQR)</td>
<td>4 (2–7)</td>
</tr>
<tr>
<td>No. of interventions during the hospital stay⁴, median (IQR)</td>
<td>2 (1–4)</td>
</tr>
<tr>
<td>Discharge discipline³, n cases (%)</td>
<td></td>
</tr>
<tr>
<td>– Medicine</td>
<td>84,327 (44.2)</td>
</tr>
<tr>
<td>– Surgery and orthopaedics</td>
<td>21,979 (11.5)</td>
</tr>
<tr>
<td>– Gynaecology and obstetrics</td>
<td>77,286 (40.5)</td>
</tr>
<tr>
<td>– Other</td>
<td>7245 (3.8)</td>
</tr>
<tr>
<td>Health outcomes, n cases (%)</td>
<td></td>
</tr>
<tr>
<td>– Unplanned hospital readmission⁵</td>
<td>15,643 (8.2)</td>
</tr>
<tr>
<td>– In-hospital all-cause death</td>
<td>3834 (2.0)</td>
</tr>
</tbody>
</table>

IQR = interquartile range * Active, secondary discharge diagnoses according to the International Classification of Diseases, 10th revision. † Diagnostic, medical and/or surgical interventions according to the Swiss classification of medical interventions (“Schweizerische Operationsklassifikation”). ‡ We stratified healthcare disciplines into four categories based on the unit from which the inpatient was discharged: medicine, surgery and orthopaedics, gynaecology and obstetrics, other disciplines (ear, nose and throat clinic, eye clinic, human genetics, pathology, or radiology). § Unplanned index hospital readmissions due to the same diagnosis as during the previous hospitalisation within 18 days after discharge are routinely recorded for inpatients in Switzerland.

Published under the copyright license “Attribution – Non-Commercial – No Derivatives 4.0”. No commercial reuse without permission. See http://enm.ch/en/services/permissions.html.
Figure 1: Multimorbidity clusters, intervention clusters and selected Bayesian networks for all-cause in-hospital death and unplanned hospital readmission (total n = 190,837 inpatient cases).

Note: For illustration purposes, subfigures (A) and (B) depict every multimorbidity and every intervention cluster, respectively, which was derived from the mutual information analysis; subfigures (C) and (D) exemplify a multimorbidity and an intervention cluster respectively, with the corresponding Bayesian network for all-cause in-hospital death and unplanned hospital readmission; subfigures (E) and (F) exemplify another multimorbidity and another intervention cluster respectively, with the corresponding Bayesian network for all-cause in-hospital death and unplanned hospital readmission. The variables ‘age’ and ‘sex’ in the Bayesian networks signify the age and sex of the inpatient cases; the corresponding variables ‘admission’ and ‘diagcount’ characterise the admission status (emergency or non-emergency admission) and the number of secondary diagnoses/comorbidities respectively.

4,1315 out of 33,942 (3.9%) patients died during the hospital stay and 3158 out of 33,942 patients (9.3%) had an unplanned hospital readmission.

Figure 1E characterises a surgical complication cluster (cluster 2) consisting of nine diagnoses. These were primarily linked via the general diagnostic node ‘incidents caused by medical measures’ (ICD-10 T84.9). Related diagnoses included unspecified surgical complications (ICD-10 T81.8), suture dehiscence (ICD-10 T81.3), postoperative haemorrhage and haematoma (ICD-10 T81.0), acute bleeding anaemia (ICD-10 D62), temporary blood clotting disorder and acquired coagulation deficiencies (ICD-10 U69.12 and D68.4), unspecified delirium (ICD-10 F05.8), and surgical site infections (ICD-10 T81.4). Overall, 25,870 out of 190,837 cases (13.6%) had at least one of these nine clustered diagnoses during the hospital stay. Interestingly – and in contrast to the unplanned readmissions outcome – we observed no direct relationship between cluster 2 and in-hospital mortality. In cluster 2, 1229 out of 25,870 (4.8%) patients died during the hospital stay and 4366 out of 25,870 patients (16.9%) had an unplanned hospital readmission.

In the last example (fig. 1F), nine cardiological interventions and diagnostic procedures – including different diagnostics and ablation techniques in cases with tachyarrhythmia, such as ablation using three-dimensional mapping (37.34.14), conventional radiofrequency ablation of tach-
and potential direct and indirect effects to be differentiated. Relationships to be easily visualised based on the concept of multi-level Bayesian networks to explore multi-level outcomes and quality measures. To our knowledge, this is one of the first studies in which Bayesian networks have been used to mine a large hospital database. In our study, the cases’ underlying characteristics (age, emergency admission status and number of comorbidities) explain the low mortality in this cluster.

Discussion

We explored the clustering of diagnoses, interventions and diagnostic procedures in a large, real-world dataset of medical and surgical inpatient cases. Moreover, we exemplified the potential use of Bayesian networks to mine healthcare databases for potential direct and/or indirect cluster effects on important health outcomes and quality measures. To our knowledge, this is one of the first studies in which Bayesian networks have been used to mine a large hospital database in order to explore inpatient complications and related high-risk clusters which may be amenable to quality improvement measures. Previous studies have applied the concept of multi-level Bayesian networks to explore morbidity at a population level [11–13]. Bayesian network analysis and consecutive directed acyclic graphs may be of particular interest, as they allow complex interrelationships to be easily visualised based on a priori knowledge, and potential direct and indirect effects to be differentiated.

Nevertheless, as with any observational analysis method, the proposed data mining technique may not allow the demonstration of definite causal relationships. Potentially relevant detrimental or beneficial effects of a given set of diagnoses and/or interventions observed in the Bayesian networks should be substantiated through specific causal investigations. In the present proof-of-concept study, we used a limited set of variables (age, sex, emergency admission status and number of secondary diagnoses), which may act – based on external, expert knowledge – as potential confounders in the specified cluster-health outcomes relationships. Further variables could be added to the Bayesian networks to examine, for instance, time trends (e.g., years, months or seasons) and to further reduce the propensity for unmeasured confounding.

Regarding in-hospital death and unplanned readmission, we observed several potential high-risk clusters which are clinically plausible, are in line with previous investigations (e.g. [14–16]), and which have not yet been selected as specific prevention targets in our hospital. A wide variety of questions could be posed, such as: Would inpatients with ablations for tachyarrhythmia or patients with certain postoperative complications (e.g., postoperative haemorrhage) benefit from a targeted quality programme to reduce unplanned hospital readmissions? Why are mechanical complications of joint endoprostheses linked to in-hospital mortality? These questions should be addressed by focused and more detailed investigations (e.g., a case-control study to evaluate risk factors for all-cause death in specific surgical populations) and subsequent quality and patients safety programmes such as the multimodal “Enhanced Recovery After Surgery” (ERAS) strategy, if appropriate. Such programmes have been shown to result in substantial quality improvements [17].

Cluster analysis combined with Bayesian network learning can be used to systematically generate hypotheses about potential high-risk clusters for a wide spectrum of health and quality outcomes. Such explorative data mining analysis may be able to provide a data-driven, objective overview of potential quality improvement targets, as many quality improvement and patient safety programmes have a narrow focus on specific patient groups, health outcomes or high-risk interventions, without any overarching, hospital-wide prevention strategy. The presented analytical strategy could be particularly important for large healthcare institutions and healthcare networks (e.g., Kaiser Permanente in the United States) with high admission numbers, allowing reasonable stratification in Bayesian network analysis. Nonetheless, it is still unclear whether an institution-wide, data-driven quality strategy is more effective and/or more cost-effective than regular, more focused quality approaches.

Our study has limitations. Firstly, we relied on routine codes and classifications, which may have led to the misclassification of diagnoses and/or interventions during the coding process. We did not have internal validation data on the inter-rater agreement for coding certain diagnoses or interventions; however, all diagnoses and interventions were coded and verified by professional medical coders using standard criteria based on a list of diagnoses and interventions, making it highly unlikely that diagnoses have been misclassified. Secondly, in our Bayesian network...
analyses we cannot exclude the possibility of unmeasured or residual confounding. Thirdly, our study results were derived from a single healthcare centre and may not be generalisable to other healthcare settings and institutions that cover different patient populations or that use other classifications to routinely code inpatient diagnoses and interventions. Lastly, we used a pragmatic approach and performed our analyses on a case level and not on a longitudinal inpatient level, which might have helped us to interpret certain relationships and signals over time [18].

As part of an explorative analysis, we identified potential high-risk multimorbidity and intervention clusters in a tertiary care inpatient population. Bayesian network analysis may be used as a tool to mine large healthcare databases in order to systematically explore intervention targets for quality improvement and patient safety programmes. However, relevant associations derived from Bayesian network analysis should be substantiated in consecutive investigations using specific causal models.

Acknowledgements
We performed all analyses at the sciCORE scientific computing centre, University of Basel, Switzerland.

Disclosure statement
No financial support and no other potential conflict of interest relevant to this article was reported.

References
Appendix 1
Description of diagnosis/multimorbidity clusters (coded according to the German modification of the International Classification of Diseases, 10th revision [ICD-10]). Available as separate PDF file at https://smw.ch/article/doi/smw.2020.20299.

Appendix 2

Appendix 3

Appendix 4