Managing Complexity. From Documentation to Knowledge Integration and Informed Decision
Findings from the Clinical Information Systems Perspective for 2018

Werner O. Hackl1*, Alexander Hoerbst2*, Section Editors for the IMIA Yearbook Section on Clinical Information Systems
1 Institute of Medical Informatics, UMIT - Private University for Health Sciences, Medical Informatics and Technology, Hall in Tirol, Austria
2 eHealth Research and Innovation Unit, UMIT - Private University for Health Sciences, Medical Informatics and Technology, Hall in Tirol, Austria
* Equal Contribution

Summary
Objective: To summarize recent research and to propose a selection of best papers published in 2018 in the field of Clinical Information Systems (CIS).

Method: Each year a systematic process is carried out to retrieve articles for the CIS section of the IMIA Yearbook of Medical Informatics and to select a set of best papers for the section. The same query as in the last five years was used. The retrieved articles were categorized in a multi-pass review carried out by the two section editors. The final selection of candidate papers was then peer-reviewed by Yearbook editors and external reviewers. Based on the review results the best papers were then chosen at the selection meeting with the IMIA Yearbook editorial board. Text mining, and term co-occurrence mapping techniques were again used to get an overview of the content of the retrieved articles.

Results: The query was carried out in mid-January 2019, yielding a consolidated, deduplicated result set of 2,264 articles which had been published in 957 different journals. This year, we nominated twelve papers as candidates and three of them were finally selected as best papers in the CIS section. Again, the content analysis of the articles revealed the broad spectrum of topics which is covered by CIS research.

Conclusions: We could observe ongoing trends from our 2017 analysis. The patient increasingly moves in the focus of the research activities and trans-institutional aggregation of data is still an important field of work. The move to use patient and other clinical data directly for the patient and to support data driven process management, the move away from clinical documentation to patient-focused knowledge generation and support of informed decision, is gaining momentum by the application of new or already known but, due to technological advances, now applicable methodological approaches.

Keywords
Medical informatics, International Medical Informatics Association, Yearbook, Clinical Information Systems

Introduction
In our synopsis for the 2018 issue of the IMIA Yearbook of Medical Informatics [1] we concluded that, modern clinical information systems served as backbone for a very complex, trans-institutional information logistics process and that data was more and more reused for multiple purposes. Last year we found a lot of examples showing the benefits and novel approaches to tackle the challenges of such data reuse. And we found that the patient was moving in the focus of interest of CIS research. This year we observed similar results.

About the Paper Selection
The selection process in the CIS section is stable now for five years. It is described in detail in [1] and the full queries are available upon request from the corresponding authors.

The queries were carried out in mid-January 2018. This year the search result set comprised 2,264 unique papers. From these papers 2,160 were retrieved from PubMed and 104 additional publications were found in Web of Science®. The resulting articles had been published in 957 different journals. Table 1 depicts the top twenty journals with the highest numbers of resulting articles. This year we used RAYYAN, an online systematic review tool [2] to carry out the multi-pass review by the two section editors (AH, WOH). Both section editors independently reviewed all 2,264 publications. In the first step, ineligible articles were excluded based on their titles and/or abstracts (section editor 1: n=2,184; section editor 2: n=2,140). In this first step, the agreement between the two editors was n=2,089 for “exclude” and n=29 for “not exclude” (i.e. include). The remaining 146 conflicts were solved on mutual consent which resulted in nine additional inclusions. The final candidate selection from the remaining 38 publications was done based on full text review and yielded 12 candidate papers for the CIS section 2019.
This list was then reviewed by the Yearbook editors who checked if any articles had also been selected for other sections. As in the last year, no overlaps were found, and all twelve candidates were sent in the peer-review process for the Yearbook. For each paper at least five independent reviews were collected. During the selection meeting held on April 26, 2019 in Paris where section editors from all Yearbook sections participated, three papers [3–5] were finally selected as best papers for the CIS section (Table 2). A content summary of these three best CIS papers can be found in the appendix of this synopsis.

**Findings and Trends:**

**Clinical Information Systems Research 2018**

During the selection of the best papers we as section editors get a broad overview on the research field of our CIS section. Of course this overview may be biased and to avoid selective perception we again applied our more formal text mining and bibliometric network visualizing approach [6] to summarize the content of titles and abstracts of the articles in our CIS result set as in the previous years [1, 7, 8].

This year we also extracted the 21,264 authors’ keywords from all articles and present their frequency in a tag cloud (cf. figure 1). We found 7,250 different key words, of which 5,149 were only used once. Most frequent keywords were “human” (n=699) followed by “female” (n=296), “male” (n=263), “electronic health record(s)” (n=235), “adult” (n=230), “child” (n=191) and “health communication” (n=164).

Figures 2 and 3 depict the resulting co-occurrence maps of the top-100 terms from the titles and of the abstracts of the 2,264 papers of the recent CIS result set. In 2017 we discovered six different clusters of terms within the titles, last year four and this year we found five main clusters. Whereas “human” was the top authors’ keyword in the articles, the analysis of the titles revealed “system” as central hub for these five clusters.

The cluster analysis of the abstracts yielded also five clusters (cf. figure 3). The red cluster on top comprises terms connected with electronic health records (EHR) and requirements, data sharing, interoperability, architecture and integration. The blue cluster on the right side is dedicated to geographic information systems and related terms. The violet cluster on the bottom and center represents medication and adverse event related research. The yellow cluster on the bottom left as well as the green cluster on the left side can be seen as containers for context factors mentioned in the abstracts.

As in the previous years, the results of these analyses demonstrate the wide variety of the CIS domain. Again, we found a multitude of well written and highly interesting articles of first-rate quality. Among these we interestingly found a high proportion...
of articles from Chinese researchers. This is also reflected in our candidate selection where half of the twelve papers come from China or involve Chinese contributors.

From a thematic perspective we identified a growing proportion of contributions showing that the potentials and possibilities due to technological advances, increased computing power, availability of a huge amount of structured data and sophisticated algorithms now can be unlocked.

The first of the best papers in the CIS section 2019 is an example for that. Cao Xiao and colleagues developed a novel deep learning model that learns distributed patient representation from the EHR data and performs prediction for the 30-day readmissions [10]. This inspiring paper is a very well-presented example of how deep learning techniques can be applied to distill new knowledge out from existing EHR data. The approach is highly innovative, and the developed model outperforms state-of-the-art approaches in readmission prediction.

An “evergreen” within the scope of CIS research is the detection of adverse events within EHR data. This year we chose a contribution by Jiebin Chu and colleagues who used neural attention networks for that [12]. Data Mining also was an important keyword during the last years in CIS research. This year we didn’t find this keyword very often (cf. figure 1). However, a very interesting article introducing a novel data or text mining technique termed as optimized swarm search-based feature selection [13] coming from Daohui Zeng and colleagues was included in the selection of candidate papers. The “newcomer” Blockchain as identified last year was also a prominent research topic in 2018. We included a contribution by Hongyu Li and colleagues on a blockchain-based data preservation system for medical data in our CIS best paper candidate pool.

Besides all benefits and possibilities, the use of massive EHR data sets can bring, we have to keep in mind that reality and data not necessarily are congruent. Without careful
consideration of the context in which these data are produced many pitfalls and sources of bias may be overlooked. Denis Agniel and colleagues, with their very worth reading candidate paper, remind us of this eternal truth. Another core aspect that may not be overlooked within the CIS domain are the users. Pantelis Natsiavas and colleagues propose a comprehensive user requirements methodology for secure and interoperable health data exchange [18]. This candidate paper is worth reading as it also provides a comprehensive set of user requirements and presents sets of barriers as well as facilitators for health IT solutions.

Last but not least we want to highlight the third of the best papers in the CIS section 2019. In their article which received top ratings from the IMIA Yearbook reviewers Johannes C Eichstaedt and colleagues show that the content shared by consenting users on Facebook can predict a future diagnosis of depression [3]. This study suggests that social media content may point clinicians to specific symptoms of depression. It sheds a new light on privacy and anonymity with regard to health information and vividly indicates that health related information can not only be retrieved from “core” medical and health data, but also from other user provided and publicly available data, such as social media content.
As every year, at the very end of our review of findings and trends for the clinical information systems section, we want to recommend a reading of this year’s survey article in the CIS section [19].

**Conclusions and Outlook**

We could observe ongoing trends from our 2017 analysis that the patient increasingly moves in the focus of the research activities and that the trans-institutional aggregation of data is still an important field of work. Common to these efforts is the creation of benefits for the patients. A key differentiator from what we observed in the last years are the methods applied to reach this goal. As already been described in the findings we

---

**Fig. 3**  Clustered co-occurrence map of the most relevant terms (top 60 percent, n=524) from the abstracts of the 2,264 papers in the 2019 CIS query result set. Only terms that were found in at least seven different papers were included in the analysis. Node size corresponds to the frequency of the terms (binary count, once per paper). Edges indicate co-occurrence and distance of nodes corresponds to the association strength of the terms within the texts (only top 1,000 of 27,021 edges are shown). Colors represent the five different clusters. The network was created with VOSviewer [9].
found a great number of papers – amongst them the best papers for 2018 – that were tackling problems using machine learning approaches combining them with relatively large and sometimes immense clinical data sets or so far uncommon data sets outside the core clinical domain. Apart from these ongoing developments we have also found inspiring work that deals with data driven management of processes and the use of blockchain technology to support data aggregation beyond institutional boundaries. The move to use patient data directly for the patient, to move away from clinical documentation to patient-focused knowledge generation and support of informed decision, is gaining momentum by the application of new or already known but, due to technological advances, now applicable methodological approaches.

Acknowledgements

We would like to acknowledge the support of Lina Soualmia, Adrien Ugon, Brigitte Seroussi, Martina Hutter and the whole Yearbook editorial team as well as the numerous reviewers in the selection process of the best papers.

References


Correspondence to

Dr. Werner O Hackl
Institute of Medical Informatics
UMIT – Private University for Health Sciences, Medical Informatics and Technology
Eduard-Wallnoefer-Zentrum 1
6060 Hall in Tirol, Austria
Tel: +43 50 8648 3806
E-mail: werner.hackl@umit.at

Dr. Alexander Hörbst
eHealth Research and Innovation Unit
UMIT – Private University for Health Sciences, Medical Informatics and Technology
Eduard-Wallnoefer-Zentrum 1
6060 Hall in Tirol, Austria
Tel: +43 50 8648 3814
E-mail: alexander.hoerbst@umit.at
Within 6 months preceding the documenta
in the EMR, the better their predictive power.

suggest that the closer in time the Facebook
using latent dirichlet allocation. The results
built a prediction model using the textual
the prevalence of the disease. The authors
at most 49%. Therefore, the authors con-
clude that such high rates of underdiagnosis
and undertreatment suggest that existing
procedures for screening and identifying
depressed patients are inadequate. The au-
thors introduce a novel approach by using
Facebook language data from a sample of
consenting patients to detect/predict de-
pression from that data. The study involved
the analysis of the Facebook posts of 114
patients that prior had a diagnosis of de-
pression. For each of the patients with such
a diagnose, five random patients without a
diagnose of depression in the same period of
time were added to the analysis to simulate
the prevalence of the disease. The authors
built a prediction model using the textual
content of the Facebook posts, post length,
frequency of posting, temporal posting
patterns, and demographics. A 10-fold
cross-validation was applied to avoid overfit-
ting. To yield interpretable and fine-grained
language variables, 200 topics were extracted
using latent dirichlet allocation. The results
suggest that the closer in time the Facebook
data are to the documentation of depression
in the EMR, the better their predictive power.
Within 6 months preceding the documenta-
tion of depression an accuracy of 0.72(AUC)
was achieved. The authors conclude that
Facebook language-based prediction models
perform similarly to screening surveys in
identifying patients with depression when
using diagnostic codes in the EMR to iden-
tify diagnoses of depression.

Parr SK, Shotwell MS, Jeffery AD, Lasko TA,
Matheny ME

Automated mapping of laboratory tests to
LOINC codes using noisy labels in a national
health record system database

J Am Med Inform Assoc 2018;25(10):1292-300

Aggregated data from multiple data-sources
can be a valuable source for different fields
of research and other domains such as public
health or the creation of clinical evidence. A
very common problem related to the use of
data from different sources is their different
coding and the use of non-standardized
terminology. An important aspect in the
assessment of patient outcomes are laboratory
findings. In this context Logical Observation
Identifiers Names and Codes (LOINC) can
be seen as an important standard. However,
mapping laboratory findings to LOINC
codes manually can be time consuming. As
electronic health records (EHR) are a rich
source of data accumulated through routine
clinical care the authors aim to develop a ma-
chine learning algorithm to automate map-
ning of unlabeled data and reclassification of
incorrect mappings within labeled data. For
this purpose, inpatient and outpatient labo-
atory data (6.6 billion laboratory results)
from 130 Veterans Affairs (VA) hospitals
was collected. The dataset used for training
contained an unknown number of labelling
errors and was not manually cleaned (noisy
labelling approach). They implemented lo-
gistic regression, a random forest multiclass
classifier, and a 1-versus-rest ensemble of
binary random forest classifiers. All models
were refined with a 5-fold cross-validation.
Although the mapping results of the models
are not in all cases convincing, the authors
investigated important reasons for this
matter. In addition, they were able to prove
that their results are similar in accuracy to
the best reported automated methods for
laboratory test mapping. This is a remarkable
result as the model was built using noisy
data. The approach is one of the first that can
be applied fully automated with no need of
manual intervention.

Xiao C, Ma T, Dieng AB, Blei DM, Wang F

Readmission prediction via deep contextual
embedding of clinical concepts


Hospital readmission is a critical figure in
the assessment of the quality of a treatment
process. Indeed, many hospital readmis-
sions are avoidable and pose a certain risk
for the patient. Although it is not an easy
task to predict hospital readmission as it is
not only related to the disease and the treat-
ment but also a complex set of risk factors
which are interrelated. The current paper
aims at presenting a hybrid deep learning
model structure that combines topic model-
ing and Recurrent Neural Network (RNN)
to distill the complex knowledge hidden
in those contexts and perform accurate
readmission prediction. The proposed
‘CONTENT’ model covers both the global
and local contexts within the patient jour-
ney from an EHR through a hybrid Topic
Recurrent Neural Network (TopicRNN)
model. It transforms patients’ complicated
event structures into deep clinical concept
embedding, which can be viewed as a novel
form of patient representation encoding
the patient clinical conditions from both
long and short terms. In order to build that
model the EHR data (including disease,
lab findings and medication codes) from a
cohort of 5,393 patients with congestive
heart failure was used. The model outputs
a context vector for each patient, which
characterizes his/her overall condition.
The proposed model outperforms existing
methods in readmission prediction (e.g.
0.6103 ± 0.0130 vs. second best 0.5998 ±
0.0124 in terms of ROC-AUC). The derived
patient representations were further utilized
for patient phenotyping. The subgroups
allow a better understanding of different
readmission risks in the cohort.