CLINICAL CONTENT COVERAGE EVALUATION ON ICD-11:
THE THIRD IN SERIES

by
Xiaohan Zhang

A thesis submitted to Johns Hopkins University in conformity with the requirements for the degree of Master of Science in Health Sciences Informatics

Baltimore, Maryland
August 2019

©2019 Xiaohan Zhang
All rights reserved
Abstract

**Background and Objective:** To better help store and render computer-based patient records in an accurate and efficient way, standardized and robust clinical classifications and taxonomies should be developed. To investigate whether major coding systems can solve the issue and help develop a reliable approach for evaluation, the authors conducted the study to look into the content coverage of major coding systems.

**Methods:** Clinical texts from four medical centers were retrieved and organized. A total of 14,247 words were parsed and 3061 clinical concepts were identified. The concepts were clustered into five different semantic groups. Each concept was then coded by ICD-11, SNOMED CT International, and Monarch Disease Ontology (MONDO). The score scaling is described here: 0 = no match, 1 = fair match, 2 = complete match.

**Results:** SNOMED CT International had an overall score of 1.55 and stood out in every category, except for Diagnoses, where ICD-11 demonstrated 1.86 in this category. ICD-11 had an overall score of 1.10 and showed improvements in every category, compared to ICD-9-CM and ICD-10-CM. The overall score MONDO is 0.43 but it scored 1.28 in Diagnoses.

**Conclusions:** No present classification systems captured all the concepts. SNOMED CT International has the highest overall score. ICD-11 performs better than SNOMED CT International in representing Diagnoses and performs better in every category than ICD-9-CM and ICD-10-CM. Considering the high score of ICD-11 in diagnoses and its fast internet-based developing environment, ICD-11 could be our next step towards improving healthcare information exchange (HIE).

**Keywords:** Electronic Health Record, clinical classifications, taxonomy, evaluation,

Primary reader and Preceptor: Christopher G. Chute
Contents

Background

EHR .................................................................................................................................1
Previous ICD Classifications ......................................................................................1
ICD-11 Design & Structure .............................................................................................4

Literature review ..............................................................................................................5

Methods

Classifications ....................................................................................................................7
Clinical Text Sources .......................................................................................................8
Scoring ............................................................................................................................10
Analysis ..........................................................................................................................12

Results

Comparison of SNOMED/ICD-11/MONDO .................................................................13
Comparison of SOMED/ICD-11/MONDO for Unique set ...............................................15
Comparison of ICD-9-CM/ICD-10-CM/ICD-11 ...........................................................16

Discussions .......................................................................................................................17
Limitations .......................................................................................................................18
Conclusions ....................................................................................................................20
Appendix .........................................................................................................................21
Reference .........................................................................................................................22
CV ....................................................................................................................................23
Background

Electronic Health Records (EHR)

Development of Information Technology (IT) has brought out the rapid growth of healthcare. Under such scenario, the wide use of electronic health records (EHRs) in the US is inevitable and foreseeable. Patients’ conditions and events along with the relevant procedures are the core of patient care. To deal with the large quantity of patient data, and to store, use and exchange the data, computer-based tools and systems are needed. To better assist, standardized coding systems and terminologies are needed in representing and interpreting clinical text into computer-friendly codes. Plenty of resources have been contributed from research filed and politics to contribute to the process of developing such coding systems. Some of these predominant systems are The Systematized Nomenclature of Medicine (SNOMED) and The International Classification of Diseases (ICD) classifications.

Previous ICD classifications

The International Classification of Diseases (ICD) is a held by World Health Organization (WHO)and designed as a tool for recording, exchanging and reporting healthcare related conditions and events. The ICDs have been fulfilling their designed purposes as interpreting and recording clinical concepts since they were developed. What is more, besides their original uses in hospitals, clinics, and research fields, they are also widely adopted in billing and administrative institutions to accommodate the needs of regulating health industries.

There are controversial arguments about the date when ICD classifications can be traced back to. But surely there were several events which contributed a lot to the development of this series of classifications. In 1853 during the first International Statistical Congress, Jacob Marc
d’Espine and William took the assignment to develop a system of classifying mortality which could be used universally across different languages. In 1860, Florence Nightingale proposed to develop the first set of systematic collections of hospital data. In 1893, Parisian statistician Jacques Bertillon established the first “International List of Causes of Death” [1]. It was later presented to International Statistical Institute in US and adopted by countries in North America. This “International List of Causes of Death” was revised in 1900, 1910, 1920, 1929 and 1938 #5#. In the first five revisions, contents for the whole classification system was contained in one book volume, including an alphabetic index and a tabular list. In 1948, WHO officially assumed the responsibility of the classification system and in the next year, another volume designed for the coding for causes of morbidity was added and marked the publication of the sixth revision. Therefore, the system was renamed from “International List of Causes of Death” to ‘The International Classification of Disease’ to reflect the structural changes. #4# With the introduction of the morbidity volume, the need for including mental conditions was fulfilled by a section on mental disorders for the first time. The seventh and eighth revisions were published in 1957 and 1968[1]. During the time ICD-7 and ICD-8 came into force and became more prominent, some countries produced national adapted versions to cater to the additional needs for application of ICD to their own situation.

In 1977, ICD-9 was published and the prominent change of this revision was its remarkable addition of four-digit subcategories and optional five-digit subdivisions. The changes happened as a response to the growing trend of utilizing ICD classifications to interpret and represent the specialists’ statistics and as a solution to some subjects which have been demonstrated inappropriate structurally before. ICD-9-Clinical Modification (CM) was published by US National Center for Health Statistics (NCHS) in United States to allow enhancement to morbidity
part and a more frequent update to the classification. The CM expansion is updated on October 1 yearly. Rather than a separate revision, ICD-9-CM is a separate set of 3 volumes with the first 2 including diagnosis codes and the 3rd including procedural codes for surgical, diagnostic and therapeutic procedures. This set was designed for better diagnostic coding of both inpatients and outpatients and for clinical use. While the procedural code part was included in the CM expansion, it was never prominent with the use of Current Procedural Terminology (CPT), which was published in 1966[1, 2] Since 1979, the United States required ICD-9-CM codes for Medicare and Medicaid claims, and the industry followed.

In 1983, work on ICD-10 started and it was endorsed by World Health Assembly in 1990 and introduced for annual review in 1992. It was first used in 1994. The revision added profound granularity to the classification with a size of 155,000 from an expansion of 17,000 codes. The adoption happened when 1 Jan 1999 the ICD-10 was announced to report mortality and ICD-9-CM was still used to report morbidity. NCHS was permitted by WHO to create modifications of ICD-10, which resulted in the ICD-10-CM for diagnosis codes to replace the first 2 volumes and ICD-10-PCS for procedural codes to replace volume 3. The adoption of ICD-10 was relatively swift with several countries having created their own adapted versions such as ICD-10-AM in Australia and ICD-10-CA in Canada. While at the same time, the adoption of ICD-10-CM, which was initially proposed by US Department of Health and Human Services (HHS) to replace ICD-10-CM in 1 October 2013, was then delayed twice and the final transition happened on 1 October 2015.
ICD-11 Design & Structure

ICD-11 has been accepted by WHO on 25 May 2019 to come into effect on 1 Jan 2022\(^3\). Its design and development have been going on an internet-based workplace open for proposal submissions of updates for ICD-11#ICD-11 website#. The proposal and review mechanism on this online platform make the whole workflow transparent and also makes it rich enough to spare need for countries to work alone on nation-specific modifications. The most innovative part of the structural change of ICD-11 is that it is based on a Foundation Component, from which multiple tabular lists can be derived.

**Table 1. ICD 11 Terminology\(^4\)**

<table>
<thead>
<tr>
<th>ICD-11 Term</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Foundation Component</td>
<td>Basic data contents of ICD-11, which includes necessary information to generating the fundamental alphabetical index and the foundation tabular list.</td>
</tr>
<tr>
<td>Linearization</td>
<td>Specialized linearization could be generated from the foundation component. Of them Morbidity and Mortality Statistics (MMS) is used most prominently.</td>
</tr>
<tr>
<td>Stem code</td>
<td>Stem codes are codes that can be used alone without more modifications. In other words, stem words are the minimum units of information in ICD-11.</td>
</tr>
<tr>
<td>Extension code</td>
<td>Extension codes are codes designed to add more information to stem codes in a standardized way to provide more information under a clinical scenario.</td>
</tr>
<tr>
<td>Precoordination</td>
<td>Stem codes may contain more information in a pre-combined fashion. <strong>Example:</strong> BD50.40 Abdominal aortic aneurysm with perforation</td>
</tr>
<tr>
<td>Cluster Coding</td>
<td>Cluster coding refer to the fashion in which multiple codes are presented together to describe a clinical concept.</td>
</tr>
<tr>
<td>Postcoordination</td>
<td>Postcoordination refers to linking multiple codes together through cluster coding to describe the concept in a standardized way.</td>
</tr>
</tbody>
</table>

The Foundation Component is structured multidimensionally for all ICD-11 entities, including diseases, disorders, injuries, external causes, signs and symptoms. The Foundation Component part also has other necessary information to build a specialized tabular list upon it to
represent clinical concepts. From Foundation Component, multiple tabular lists can be generated to represent concepts in different contexts, and these tabular lists are so-called ‘congruent tabular lists’ or ‘linearization’. They include: Mortality and Morbidity Statistics (MMS), Primary Care Low Resources Settings (PCL), Verbal Autopsy (VA), Startup Mortality List (SMoL). The full name of these linearization lists will always include ‘ICD-11’, e.g. ICD-11 MMS, which is used widely to report statistics about morbidity and mortality among the implementors.

In a generated tabular list, different from their previous structured patterns in Foundation Component, entities become categories, in a way where they are mutually exclusive and jointly exhaustive and linked to a mono hierarchical tree, which means the downstream entities will have only one parent, but multiple presents are still embedded in Foundation Component if available.

The innovations and rich and profound contributions made from about 300 specialists of over 50 countries have drawn great expectations to ICD-11. Several detailed external reviews of ICD-11 have been finished in the past to note its ongoing progress. The description logic behind ICD-11 also makes it grow from a classification system or hierarchical taxonomy into an ontology, which computers can derive meaning by traversing the various relationships. Yet there have been no specific studies from academics to test the functionality of ICD-11 in real clinical data. We consider our study to be the first to test the clinical coverage of ICD-11 in patient data sourced from real world electronic health records.

**Literature Review**
A comprehensive literature review was carried on to look into the scope of the research question. The search was done in Medline and Embase to exhaust the search results and cover as many research outputs as possible. Medical Subject Heading (MeSH) and keywords of ‘taxonomy/ontologies/ICD classification’, ‘evaluation/assess/test’, “electronic health record/electronic medical record/’ were included. The results were examined and filtered by researchers. Some literature is also included subjectively by researchers. The Study would be included if their topics involved tests and evaluations of the classifications/taxonomies/ontologies themselves on their performances in interpreting and presenting clinical concepts. A total of 190 papers were retrieved from Medline and 249 were retrieved from Embase. After duplicates were removed, 390 articles were examined by researchers. Among them, 35 papers were excluded for being unavailable in any database and eventually 17 were tagged as relevant for demonstrating the topic.

It is demonstrated the needs and attempts for a specific coding systems have been done in many fields, include dental, nursing, genomic, pharmaceutical and clinical research[5-9]. Monsen et.al looked into how coding systems help along with respect to the documentation of social determinants present in EHR, which is an important entity in EHR systems to provide enough details and annotations about the patients[10]. Fung et.al explored a combined way of using standardized coding systems and whether the combination could help prove the performance[11]. While there is still a gap for research about evaluation through different coding systems and an efficient evaluation matrix is expected with emerging coding systems. The limited pool also proved it high potential to find robust ways to evaluate the predominant coding systems in the US nowadays.
Methods

Classifications

This third in series evaluation inherits the logic of its precursors. The evaluation is conducted in the scope of publicly available and reliable coding systems that are structured to cover clinical events and conditions involved in patient care and its corresponding electronic health records. Table 1 includes the classifications, their sources, versions and release dates. All ontological codes were retrieved and applied in machine-friendly form and reflected updates till July 2019. A brief summary on availability and nuanced purposes on the coding systems is depicted in Appendix. ICD 11 and Monarch Disease Ontology (MONDO) are experiencing long lasting updates during the period of the analysis, but SNOMED CT International is more mature at this stage. To make longitudinal comparisons between ICD classifications, previous studies and coding systems (ICD-9-CM, ICD-10-CM) were also included[12].

Table 2

Coding Systems and Sources

<table>
<thead>
<tr>
<th>Source</th>
<th>Version</th>
<th>Release Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICD-11</td>
<td>WHO</td>
<td>04/2019</td>
</tr>
<tr>
<td>SNOMED CT</td>
<td>College of American Pathologists</td>
<td>International</td>
</tr>
<tr>
<td>MONDO</td>
<td>Oregon Health &amp; Science University et al.</td>
<td>2018</td>
</tr>
</tbody>
</table>
Clinical Text Sources

Clinical texts were randomly retrieved from four practices: Mayo Clinic (Rochester, MN), Kaiser Permanente (northern California), Park Nicollet Medical Center (Minneapolis, MN) and University of Nebraska Medical Center (Omaha, NE). For each patient involved in the study, four samples of clinical texts were retrieved from each source in the scope of outpatient, inpatient, and nursing progress notes, and radiology, operatic and consultant reports. A balanced age and gender distribution was strained for the study.

From the pooled clinical texts, 50 notes were sampled by previous researchers to yield a total of 3061 concepts to maximize the reliability of the analysis. The notes were previously categorized into two sets: Set 1 (995) includes patient history and physical dictations (341), nursing notes (94), consulting notes (161), outpatient progress notes (251), and inpatient progress notes (148); Set 2 (2066) includes discharge summaries (534), outpatient progress notes (862), radiology reports (410), and operative reports (260). 14247 clinically meaningful phrases/words were identified from the parsed samples. This two-set design was originally to test and argue about the replicability of the study. And we deliberately ignore this two-set design in this study.

Texts were subjectively by the principle researcher (CGC) into concepts and a semantic type was assigned to each of them at the same time in these series twenty years ago. The semantic types included ‘Diagnoses’ for terms labeled as ‘Diagnosis’ and ‘Primary Diagnosis’; ‘Findings’ for terms labeled as ‘diagnostic examination’, ‘physical findings’, ‘symptoms’ and ‘functional status’; “Other” for terms labeled as ‘administrative’, ‘demographic’, ‘disposition’, ‘education’,

To maintain the authenticity and reliability of the evaluation, the assigned semantic types were rechecked and turned out to remain unchanged. An example of texts is referenced here:

(18 yr. old female) Underwent biopsy of a. left thigh mole. This mole had been present for approximately 2 years but it recently changed with some central clearing.

This was removed by her local physician and at that time of its removal it measured 0.9 X 0.8 cm. On pathology, which was reviewed here, it was identified as superficial spreading melanoma Clark’s level 2, depth of invasion 0.84 mm.

The sample texts yielded these concepts and the semantic types were tagged along:

(Demographic): 18 yr

(Demographic): female

(Diagnosis): melanoma

(Extent): Clark’s level 2

(Quantitative): 0.84 mm [depth of invasion]

(Mode): superficial spreading
Still, the quantitative measures like the values of lab measures and sizes of the neoplasms are ignored deliberately. But the invasion depth is highly associated with clinical staging and severity of such diseases like melanoma, we decided to include them in our final concept list. Overall, after processing and filtering, a total of 3061 concepts were identified. The assigned semantic types and semantic groups were checked by previous researchers and re-checked in this study.

**Scoring**

A primary reviewer was assigned for MONDO and SNOMED, and two primary reviewers were assigned to ICD 11. The primary reviewers were responsible for identifying relevant codes from their assigned coding system and judging if the codes would capture the meaning of the clinical concepts. The identified codes and their descriptive texts are retrieved and entered into the file for the secondary reviewers to check. A score was given by reviewers subjectively for each concept: 0 if there is no match in the corresponding coding system; 1 for approximate matches; and 2 for a complete match. Synonyms are acceptable based on the reviewers’ judgement. If a code in the coding system could represent multiple identified concepts, then this is considered reasonable to use this code as evidence for scoring of multiple concepts.

An example of SNOMED coding system is:
18 yr. old [age or date of birth] Demographic

female [sex or gender] Demographic

melanoma Primary Diagnosis

Clark's level 2 Extent

0.84 mm [depth of invasion] Quantitative

superficial spreading Severity

thigh Anatomy

left Topology

Codes:

59223006 Juvenile

1086007 Female structure

254730000 Superficial spreading malignant melanoma of skin

50542000 Clark melanoma level 2

371305003 Skin structure of thigh

7771000 Left

17456000 Breslow measurement - depth from 0.76 to 1.75 mm

An example of ICD 11 coding system is:

18 yr. old [age or date of birth] Demographic

female [sex or gender] Demographic
2 melanoma  Primary Diagnosis

0 Clark's level 2  Extent

0 0.84 mm [depth of invasion]  Quantitative

2 superficial spreading  Severity

2 thigh  Anatomy

2 left  Topology

Codes:

XT6S  Young adult

XK8G  Left

XA5S78  Thigh

XH08X7  Superficial spreading melanoma

2C30.0  Superficial spreading melanoma, primary

After the initial scoring file was finished by the primary reviewer, the scoring was delivered to a second reviewer to independently review and validate primary scorer’s work. Changes were discussed between the primary and secondary reviewers for that specific coding system. And a consensus was achieved after the discussion. Final assignments were sent to one of us (XZ) for analysis.

Analysis

The scores were organized and clustered by its assigned semantic group and an overall score is calculated through that specific coding system alone. An average score was calculated for each of the coding systems. Considering the fact we deliberately ignored the initial two-set
design of the previous studies, we did not compare the situation of the two sets. We used R (R Core Team, 2017) to perform the analysis\textsuperscript{[13]}.

Same situation with redundant representations in the same patient records still resides. Initially, the replicates within one patient were deleted. But we decided to follow the same logic where we preserved the replicates across different patient records, such as 60 duplicates of “right” and multiple duplicates of ‘male’ or ‘female’. To examine if these reserved duplicates have introduced biases into our analysis, a definition of ‘unique set’ was introduced and applied the analysis.

Table 3

Score Distributions and Averages for Full Sample

<table>
<thead>
<tr>
<th></th>
<th>Scores</th>
<th>Score</th>
<th>Scores</th>
<th>Score</th>
<th>Scores</th>
<th>Score</th>
<th>Scores</th>
<th>Score</th>
<th>Scores</th>
<th>Score</th>
<th>Scores</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0 1 2</td>
<td>Avg</td>
<td>0 1 2</td>
<td>Avg</td>
<td>0 1 2</td>
<td>Avg</td>
<td>0 1 2</td>
<td>Avg</td>
<td>0 1 2</td>
<td>Avg</td>
<td>0 1 2</td>
<td>Avg</td>
</tr>
<tr>
<td>ICD-11</td>
<td>2 9 8</td>
<td>1.86</td>
<td>27 30</td>
<td>43 1.17</td>
<td>40 8 32</td>
<td>1.11</td>
<td>47 48</td>
<td>5 0.58</td>
<td>70 9 21</td>
<td>0.52</td>
<td>37 17</td>
<td>46 1.1</td>
</tr>
<tr>
<td>SNOMED</td>
<td>8 11 3</td>
<td>1.73</td>
<td>9 14 7</td>
<td>1.67</td>
<td>18 9 73</td>
<td>1.55</td>
<td>14 28 58</td>
<td>1.43</td>
<td>33 13 55</td>
<td>1.22</td>
<td>16 13 71</td>
<td>1.55</td>
</tr>
<tr>
<td>MONDO</td>
<td>5 0 12</td>
<td>0.28</td>
<td>8 9 3</td>
<td>0.55</td>
<td>84 2 14</td>
<td>0.29</td>
<td>85 1 14</td>
<td>0.28</td>
<td>100 0 0</td>
<td>0.76</td>
<td>4 19</td>
<td>0.43</td>
</tr>
</tbody>
</table>

Result

Comparison of SNOMED/ICD-11/MONDO

There are significant differences among the included three coding systems with regard to their abilities to capture clinical concepts. The subjective 0 – 2 integer scaling of the three included
coding systems are depicted in Figure 1. The detailed scoring and averages for these systems are presented in Table 3.

ICD-11 scored the best in diagnosis, with nearly ninety percent coverage for all diagnosis and primary diagnosis concepts captured in the clinical texts. Considering the ICD classifications are used predominantly in clinical fields and its promising future of continuing to contribute in clinical settings, the coverage rate is considered to be of high performance. The average for modifiers are good (1.11), especially when comparing it to its precursors of ICD-9-CM and ICD-10-CM. The reasons behind the scene are due to the enrichment of terms of topology and more detailed anatomical terms. Considering its fast-developing status of the foundation component, it remains promising to see if the score can be improved soon.

![Score Averages by Coding Systems](image)

**Figure 1** Scores by semantic groups for coding systems. Bar graphs of mean scores over all concepts within semantic group and overall. The 0,1,2 scaling reflects a subjective evaluation of an captured concept, 0 = absent, 1 = partly captured, 2 = complete match. Classifications depicted are from ICD-11, MONDO and SNOMED CT International. Additional analysis between ICD classifications are depicted below.
SNOMED CT International stands out in most semantic groups and overall score, and its score in diagnoses group is relatively close to that of ICD-11. The scores are relatively close between groups, which suggests its stable performances in capturing vast clinical concepts.

MONDO Ontology has the least score in all categories. But notably, the score in ‘diagnoses’ group is significantly higher than other groups, it caters to the nature of its design as ‘aims to harmonize disease definitions across the world’, and its advantage of 1:1 equivalence axioms connecting to other resources are not presented in our analysis. Considering the relatively short time of its development, there is still space for MONDO to improve in those low-score categories.

**Comparison of SOMED/ICD-11/MONDO for Unique set**

Comparison of score distribution across the three included coding systems is done between ‘Actual’ set and ‘Unique’ set, which removed all the redundant concepts (e.g. if concepts like ‘right’ showed more than once in the captured concepts, then the duplicates would be removed). The three included coding systems (ICD-11, SNOMED, MONDO) do not show statistically significant differences between ‘Actual’ set and

<table>
<thead>
<tr>
<th>Semantic Group</th>
<th>SNOMED CT</th>
<th>ICD-11</th>
<th>MONDO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diagnoses</td>
<td>1.73 (1.71)</td>
<td>1.74 (1.71)</td>
<td>1.28 (1.19)</td>
</tr>
<tr>
<td>Findings</td>
<td>1.33 (1.15)</td>
<td>1.31 (1.15)</td>
<td>0.25 (0.19)</td>
</tr>
<tr>
<td>Modifiers</td>
<td>0.29 (0.24)</td>
<td>0.28 (0.24)</td>
<td>0.09 (0.07)</td>
</tr>
</tbody>
</table>

‘Unique’ set within each semantic group. All coding systems did not show significant differences in ‘diagnoses’ and ‘findings’ groups, while ICD-11 showed lower score in unique set of
‘Modifiers’, which suggests our included concepts might be biased by the differences between these two sets. But the higher score in ‘actual’ set also suggests the coding system performs better on more frequent concepts, which proves the usability of the coding system in real settings.

**Comparison of ICD-9-CM/ICD-10-CM/ICD-11**

![Score Average by ICD Classifications](image)

*Figure 2* Scores for ICD classifications (ICD-9-CM, ICD-10-CM, ICD-11), broken down by semantic groups and overall score. The scaling is the same as in this study.

The score for ICD-11 was retrieved and analyzed with data from previous studies to make a comparison across the ICD classifications\[14\]. All ICD classifications performed best in ‘Diagnoses’ group than in other groups. And ICD-11 showed a remarkable improvement even in this group, when a score of 1.86 left it with little space to improve. ICD-11 also performed better in other semantic groups as well as in overall score. There are multiple reasons behind the significant improvement: 1) Introduction of new terms including more detailed topology and anatomical terms enables the classification to depict relevant concepts; 2) Systems of post
coordination and cluster coding add up to the capacity of ICD-11 to describe more complex clinical concepts; 3) New sections of ICD-11 like ‘Extension’ codes remarkably improved the performance, which could be told when scorers were doing the analyses. However, ICD-11 didn’t improve evidently in ‘Treatment and Therapeutic Procedures’ and ‘Other’ groups, which suggests its potential and need of more input in codes for these parts. A potential clinical modification on ICD-11 about the procedures and a potential combination of ICD-11 and drug-oriented coding system may improve the performance.

Discussion

The study is the third in series to evaluate the content coverage for major clinical classification systems. Overall, the major systems these days still fail to interpret the captured clinical concepts. The capacity and validity to support clinical, academic projects and coordinate billing activities seem insufficient. More output should be contributed to the field to improve these systems.

Notably, ICD-11 has significant improvement when compared to ICD-9-CM and ICD-10-CM, where it is left with little space to prove in ‘diagnoses’ part and improved significantly in ‘modifiers’, which are highly relevant for describing detailed clinical cases and events. The ICD-9-CM and ICD-10-CM are using widely in academic, clinical and business fields. The developing complexity of the patient cases and increasing abundancy of medical terms have made it more necessary to have reliable and accurate clinical classification systems. Many research outputs based their methodology on ICD classifications to do baseline processing and analysis, like using ICD codes to find a group of patients with a specific phenotype. The methodology could also be
hybrid, but ICD codes are quite frequently used these days. Different specialized ontologies like MONDO would also make it easier for computer science researchers to contribute to clinical world. Our serialized analyses on major coding systems can provide concrete evaluation of these included systems about their functions in clinical concept representation and interpretation across many healthcare domains.

Limitation

The present study still has many limitations. The corpus of the included clinical concepts are relatively small. The material to generate the list only cover a small panel of patients, with the limited diagnoses corpus, the heterogeneity would also be not feasible to test. Compared with the population in the US, the panel size is too small. And even from the same material about this small group of patients, the concepts are parsed and defined subjectively, and the results may differ if we have different researchers to define the concepts. As illustrated and analyzed in the study, biases might be introduced by the redundancy of concepts across the corpus. The redundancy of the concepts within each patient and across the whole sample, could make the estimates on biases extremely hard to tell.

To be aligned with our previous studies done with ICD-9-CM and ICD-10-CM, we used the same set of concepts. Although the concept list was checked manually by one of our team members (XTZ), only the validity of the concepts are checked about if they are correctly kept there. But if the codes were not valid these days, e.g. the diagnoses hierarchies may differ from what they were twenty years ago, we would not make changes to the terms. Thus, a more compressive
way to examine whether the concepts are still fit and valid today is a hard but important issue for the study.

The 0-2 scaling may lack granularity for fully describing concepts. It can be easy to arbitrarily assign 0 or 2 to a concept. But when evaluating some concept which is not explicitly recorded in that coding system, assigning a ‘1’ would be the only option but its rationality remains to be explored, since ‘0, 1, 2’ would perform better to show the rank but may not be that sufficient when arguing about the overall robustness of that specific coding system. In this way, the calculations based on average score may need further discussion.

The lack of a systematic way to do the scoring may decrease the authenticity of the final scoring. As researchers subjectively gave the score, the score itself needs to be validated and the approach we discussed the mismatch and came up with a consensus may not be robust.

Similarly, as in our previous studies, we are still measuring a rapidly developing target. Though SNOMED CT International is relatively stable, with at least yearly update itself, the other two systems are experiencing frequent updates over the time we did the analysis. The situation even became more significant when ICD-11 applies an online environment for proposal discussion and approval. The highly instant and interactive environment gives high potential to move forward and improve, but it also triggers a case when instability makes the system extremely hard to evaluate.
Conclusion

We conclude that still no present coding systems could capture clinical concepts depicting patients’ situation to a satisfying degree. But ICD-11 showed significant improvement, compared to its precursors from ICD classifications (ICD-9-CM/ICD-10-CM). We should bring more resources onto the evaluation of coding systems and target to develop a robust and standardized approach for evaluation to gibe the public accurate and convincing measures of these predominant coding systems.
Appendix

Detailed Score Distribution by Semantic Types for ICD-11/SNOMED/MONDO
Reference

PROFILE
Master of Science candidate concentrating in informatics and clinical medicine. Bachelor of Medicine with one and a half year of clinical rotation in Chinese top hospitals. Researcher with experience in human immunotherapy and informatics. Standing at different positions in healthcare field and finally choosing to serve as a pediatrician and informatician in the future. Fluent in English and mandarin.

EDUCATION

Master of Science in Health Sciences Informatics  
*Johns Hopkins University School of Medicine - Baltimore, Maryland*  
Aug. 2017 – Present

Bachelor of Medicine in The Five-Year Program of Clinical Medicine  
*West China School of Medicine, Sichuan University - Chengdu, China*  

HONORS & AWARDS

<table>
<thead>
<tr>
<th>Award</th>
<th>Institution</th>
<th>Years</th>
</tr>
</thead>
<tbody>
<tr>
<td>Individual Award, Second Class</td>
<td>West China School of Medicine, Sichuan University</td>
<td>2013-2014</td>
</tr>
<tr>
<td>Outstanding Volunteer of Sichuan University, Sichuan University</td>
<td>2012-2013</td>
<td></td>
</tr>
</tbody>
</table>

RESEARCH EXPERIENCE & CLINICAL EXPERIENCE

Sr. Research Fellow  
*Johns Hopkins Armstrong Institute - Baltimore, MD*  
*Apr. 2018 – Present*  
- Focused research on patient safety and quality improvement
- Statistical analysis in pediatric trauma data
- Develop informatics studies in the team

Research Fellow  
*Johns Hopkins School of Medicine - Baltimore, MD*  
*Dec. 2017 – Present*  
- Focused research on FHIR resources modeling to represent clinical data
- Mapping across different oncology systems
- Training new team members

Clinical Rotation  
*2017*  
*West China Hospital - Chengdu, China*  
*Mar. 2016 - June*  
- Assisting for inpatients care during rotations.
- Participated in clinical journal presentation and case report of emergency and critical patients.
- Performed clinical operations for therapy like intubation. Served as assistant for surgeries.

Research Member  
*Department of Molecular Genetics and Microbiology in University of Florida - Gainesville, Fl*  
- Evaluated and optimized protocol for blotting in the lab
- Found convincing evidence to prove beneficial impacts of viruses on human T cells and supported CAR-T therapy
- Trained new-enrolled lab mates

INFORMATICS EXPERIENCE/ACTIVITIES

Informatics Project - Clinical Profiles  
*Dec. 2017 - Present*
Johns Hopkins University

- Mentored by Dr. Christopher G. Chute
- Define clinical profile contents for asthma patients and in charge of interoperable content translation
- FHIR-Friendly JSON model building

Informatics Project - Proposal for a Tuberculosis Real - Time Surveillance System  
Johns Hopkins University  
Mar. 2018 - May. 2018

Informatics Project - CDS tool to minimize the amount of packed red blood cells given during cardiac surgery at  
Johns Hopkins Hospital  
Mar. 2018 - May. 2018

Informatics Project - Using Stack Model to Critique the MetroHealth of Cleveland  
Johns Hopkins University  

Informatics Project - Osteoporosis Management System Proposal  
Johns Hopkins University  
Nov. 2017 - Jan. 2018

- Acted as system designer with primary contributions to informatics standards involved
- Conducted functional/non-functional requirement analysis of the system

Independent Study - De-identification and Re-identification of Patient Data  
Johns Hopkins University  
Apr. 2018 - July. 2018

- Mentored by Prof. Edward Bunker
- Exploration of efficient de-identification process of patient data

Informatics Project - ICD 11 Ontology coverage detection with Mondo and SNOMED CT  
Johns Hopkins School of Medicine  
Sep. 2018 - Present

- Mentored by Dr. Christopher G. Chute
- Third analysis in series to test ICD ontology coverage
- Proposing a research paper to report the findings

Informatics Project - Clinical Workflow Mining for Pediatric Trauma Patients  
Johns Hopkins Armstrong Institute  
Sep. 2018 - Present

- Mentored by Dr. Ayse Gurses
- Clinical log mining for pediatric trauma patients
- Potential clinical decision support software prototype designing

Independent Study - Telemedicine Applications for Real Clinical Settings  
Johns Hopkins University & Sibley Memorial Hospital  

- Mentored by Dr. Ashwini Davison
- Utilizing Alexa devices to simulate human-to-human talks to provide guides for patients
- Proposing a case report to analyze pros and cons for telemedicine techs in real clinical settings

Independent Study and Capstone Project- R - Health Direct Primary Care Project  
Johns Hopkins School of Medicine  
Jan. 2019 – Aug 2019

- Mentored by Dr. George Kim and Dr. Marion Ball

TEACHING EXPERIENCE

Health Information Systems: Design to Deployment  
Johns Hopkins School of Public Health  
Sep. 2018 - Nov. 2018

- Taught by Dr. Bob Miller
- Assisting as teaching assistant to help organize live talks and provide evaluation and feedbacks for students
- Giving detailed evaluation for students’ projects

Clinical Informatics Course for Medical Students  
Johns Hopkins School of Medicine  
Dec. 2017

- Taught by Dr. Ashwini Davison
- Assisting as teaching assistant to organize lectures and provide evaluation and feedbacks for students
- Giving Initial design ideas on course materials
VOLUNTEER EXPERIENCE

Annual Meeting of Chinese Society of Anesthesia - Chengdu, China  Oct. 2014
Promotion for Sichuan University - Xuzhou, China  Jun. 2013
West China Women and Children’s Hospital & Listening Center in West China Hospital - Chengdu, China  2012 - 2013

SOCIAL ACTIVITIES

Academic Exchange on International Medical Emergency Relief 2014 and Trauma Emergency Training Seminar  
Chengdu, China  Jul. 2013/ Jul. 2014
- Served in the interpretation team for all lectures
- Served as teaching assistant during Emergency Room simulation sections
- Coordinated tutors with the members and solved daily problems

4th International Universities Rowing Regatta - Xinjin, China  Aug. 2013
- Helped deal with emergency and offer medical consultation
- Performed clinical operations and organize discussion for members attending the seminar

PRESENTATIONS

Reusable clinical profiles with temporal patterns for machine learning applications
Kenneth Roe, PhD; Xiaohan Zhang, MD; Brant Chee, PhD; Vidhu Jawa, BE; Jordan Matelski, BS; Jeremy Epstein, MD; Richard Zhu, MD; Christopher G. Chute; MD DrPh; and Casey Overby Taylor, PhD  
Second NorthEast Computation Health Summit - Boston, Ma  Apr. 2018

SELF STUDY CERTIFICATES

Cousera  Oct 3, 2017-present
- Data Science Track - Johns Hopkins University (Accomplished with certificates acquired)
  Ten courses mainly cover R programming and data analysis
- Python for Everyone Track - University of Michigan (Accomplished with certificates acquired)
  Five courses cover Python Programming
- From Excel to MySQL Track - Duke University (Accomplished with certificates acquired)
  Five courses cover Excel and SQL skills
- Algorithms - Princeton University (Pending)
  Basics help think during programming

Udacity  Apr 10, 2018 - Present
- Natural Language Processing Nanodegree Program

SKILLS

- R Programming
- SQL
- Python Programming
- FHIR Modeling
- Photoshop
- ArcGIS
- NVivo