entropy

MDPI

Article

Leveraging Shannon Entropy to Validate the Transition between ICD-10 and ICD-11

Donghua Chen [1], Runtong Zhang [1,*] and Xiaomin Zhu [2]

[1] School of Economics and Management, Beijing Jiaotong University, Beijing 100044, China; 15113181@bjtu.edu.cn

[2] School of Mechanical, Electronic and Control Engineering, Beijing Jiaotong University, Beijing 100044, China; xmzhu@bjtu.edu.cn

* Correspondence: rtzhang@bjtu.edu.cn; Tel.: +86-010-51683854

Received: 13 September 2018; Accepted: 2 October 2018; Published: 8 October 2018

Abstract: This study aimed to propose a mapping framework with entropy-based metrics for validating the effectiveness of the transition between International Classification of Diseases 10th revision (ICD-10)-coded datasets and a new context of ICD-11. Firstly, we used tabular lists and mapping tables of ICD-11 to establish the framework. Then, we leveraged Shannon entropy to propose validation methods to evaluate information changes during the transition from the perspectives of single-code, single-disease, and multiple-disease datasets. Novel metrics, namely, standardizing rate (SR), uncertainty rate (UR), and information gain (IG), were proposed for the validation. Finally, validation results from an ICD-10-coded dataset with 377,589 records indicated that the proposed metrics reduced the complexity of transition evaluation. The results with the SR in the transition indicated that approximately 60% of the ICD-10 codes in the dataset were unable to map the codes to standard ICD-10 codes released by WHO. The validation results with the UR provided 86.21% of the precise mapping. Validation results of the IG in the dataset, before and after the transition, indicated that approximately 57% of the records tended to increase uncertainty when mapped from ICD-10 to ICD-11. The new features of ICD-11 involved in the transition can promote a reliable and effective mapping between two coding systems.

Keywords: ICD-11; ICD-10; Shannon entropy; validation; transition

1. Introduction

The International Classification of Diseases (ICD) is a global standard for diagnostic health information [1]. The ICD developed by the World Health Organization (WHO) enables sustainable and systematic recording, analysis, interpretation, and comparison of mortality and morbidity rates of different countries at various time points. Over the past 20 years, the 10th revision of the ICD (ICD-10) has been widely utilized in classifying healthcare information. The 11th revision of the ICD (ICD-11) was formally released on 18 June 2018 for testing and implementation, in accordance with specific timelines and requirements of different countries. The development of the new ICD standards will revolutionize the global medical informatics with opportunities and challenges for quality and safety in the next several decades [2].

The ICD standards have been used in medicine and healthcare for over a century. The first ICD standards initially focused on the statistics of causes of death. In 1946, the Interim Commission of the WHO took over the revision of the ICD and introduced a method for disease classification. At present, the most widely used version of the ICD is ICD-10, which was proposed in 1989. In contrast to ICD-10, ICD-11 is established upon ontology models. Some of the value sets in ICD-11 are derived from external ontologies, such as the Systematized Nomenclature of Medicine—Clinical Terms (SNOMED CT) [3].

The WHO released ICD-11 in 2018 to provide an international standard for disease classification in the 21st century [4]. Subsequently, ICD-11 will be submitted to the 144th Executive Board Meeting in January 2019, and then to the 72nd World Health Assembly in May 2019. The member states of the WHO will begin reporting the use of ICD-11 on 1 January 2022, after the endorsement.

The structure and design of the newly proposed ICD-11 are based on clinical practices over the past few decades and differ considerably from those of the previous versions of the ICD [5,6]. For example, ICD-11 provides novel concepts that allow multiple purposes for disease classification. Moreover, the code scheme that contains stem and extension codes in the ICD-11 is also new to users of ICD-10. Many challenges in the transition from ICD-10 to ICD-11 exist, such as previous practices of building a bridge between the ICD-9 content model (CM) and ICD-10-CM, due to the complexity of the structure and guidelines in ontology-based ICD-11 [7]. In addition, the lack of proper metrics on validating information changes in a different context of ICD standards also prevents us from utilizing new features of ICD-11 from the perspective of using massive ICD-10-coded data [8,9]. Therefore, the utilization of existing ICD-10-coded data in adapting new features of ICD-11 for global service needs in healthcare is essential [10].

The transition from ICD-10 to the new features of ICD-11 complicates further development of support tools for medical information systems. The foundation component (FC) [11] and the CM [12] are key components according to the reference guidance of the ICD-11. The FC is a multidimensional collection of all ICD entities in ICD-11. The entities cover diseases, disorders, injuries, external causes, signs, and symptoms. Some entities may be broad (for example, "injury of the arm"), whereas others may be highly detailed (such as "laceration of the skin of the thumb"). The CM defined by 13 attributes provides background knowledge of each ICD entity to allow for computerization. Generally, ICD-10 is organized in a tree-type structure with stems and branches. The structure details disease classification in healthcare layer by layer, until the layer reaches a unique disease code. However, the ICD-11 can be utilized for multiple parenting of diseases. For example, type-2 diabetes mellitus in ICD-11 belongs to different types of endocrine diseases and startup mortality list. A formal concept analysis in ICD-11 will greatly influence various types of medical and health data in the future [13]. In summary, reducing the gap during the transition from ICD-10 to ICD-11 is important to utilize existing ICD-10-coded data [14,15].

The rest of this paper proposes an ICD mapping framework, which facilitates transition between ICD-10-coded datasets and ICD-11-coded datasets in Section 2. Three metrics based on the Shannon entropy theory [16] to evaluate information changes during the transition are proposed. Then, we utilize an ICD-10-coded dataset to examine the performance of the method in terms of transforming ICD-10 codes to ICD-11 codes, through the proposed metrics in the process of adapting new features of ICD-11. Finally, we discuss and conclude the work. The purpose of this study is to help researchers to perform data-analysis with medical data from different ICD systems, which is very useful in enhancing clinical decision support based on existing ICD-coded data in the new standards.

2. Materials and Methods

2.1. ICD Mapping Framework

A framework that can ensure the effectiveness of the transition between different coding systems is essential [17]. For example, mapped ICD codes are also reused to ensure high-quality analysis, such as the investigation of the causes of death [18]. Our ICD mapping framework was established on ICD-related tables from ICD-10 and ICD-11, as shown in Figure 1. The ICD-10 tables include chapter, section, and code tables, whereas those of ICD-11 include simple tabulation and corresponding ICD-10-to-ICD-11 mapping tables. In addition, some ICD-10-coded datasets from our cooperative hospitals were collected for testing. The detailed roles of the datasets in the framework are as follows.

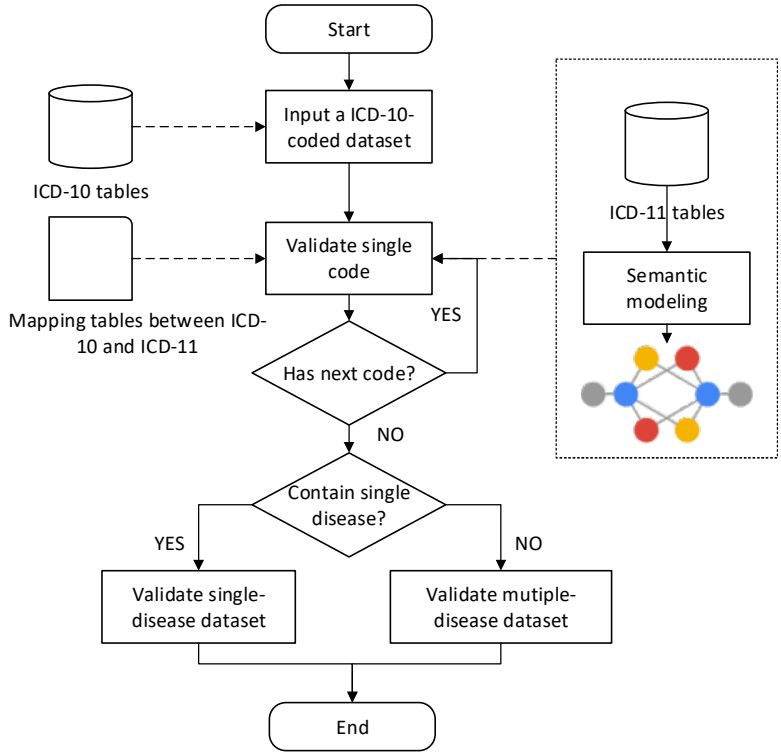

Figure 1. Overview of an International Classification of Diseases (ICD) mapping framework.

First, the ICD-11 released from the WHO in 2018 provides a simple tabulation with an easy-to-access structure to obtain relevant information of parents and children of specific ICD codes. Therefore, we construct a database to support rapid query of relevant information for the framework. In addition, a mapping table between ICD-10 and ICD-11, also acquired from the release of ICD-11, can be used to obtain one-to-one and one-to-multiple mapping relationships from ICD-10 to ICD-11 codes.

Second, the datasets of ICD-10 provide the basis of initial validation on ICD-10-coded datasets. The ICD-10-coded datasets were initially evaluated using the original ICD-10 tables to provide a benchmark for future validation, such as validating 30-day mortality across ICD-9 and ICD-10 [19]. Moreover, the tables in the ICD-10 provide comparison between standardized terms and the terms in the datasets in the context of ICD-10.

Lastly, ICD-10-coded datasets were prepared to validate the transition during mapping. Each ICD-10-coded dataset contained at least two columns, namely, textual diagnosis information and the corresponding ICD-10 code. The information in the dataset varied depending on the digital environments and use of different versions in various countries.

In summary, the aforementioned datasets from ICD-10 and ICD-11 provided the basis of the framework used in the validation between codes. A modification of the datasets was necessary before transition. The framework mapped ICD-10 codes to ICD-11 codes sequentially from code, single-disease dataset, and multiple-disease dataset levels.

2.2. ICD Validation Methods

The validation of ICD-10-code datasets during ICD mapping aims to examine legitimacy of the dataset for secondary use in ICD-11. Based on the aforementioned framework, we introduced the Shannon entropy into the process of validating information changes during the transition.

2.2.1. Single-Code Validation (SV)

The SV during mapping between ICD-10 and ICD-11 verified the changes of information between ICD codes. Three entropy-based metrics applied in the SV were the standardizing rate (*SR*), uncertainty

rate (*UR*), and information gain (*IG*) among codes. Developing such metrics and automated tools is useful for future use of consistent discovery of disease and financial analyses [20].

Validation with the *SR* refers to the examination of various ICD-10-coded datasets from different digital environments, to evaluate the difference of codes in transforming the datasets into standardized ones [21]. Generally, ICD-10-coded datasets vary from different hospitals and countries due to the consideration of quality and safety [22]. In code standardization, the mapping tables that preserve the mapping relationship, such as $c_{10} \rightarrow \{c_{11}(i) \mid 0 \leq i \leq N_{11}\}$, where N_{11} is the number of optional ICD-11 codes in the tables, are used. On the basis of the tables and a specific ICD-10 code, a corresponding ICD-10 code with the longest length of code in the mapping table provided by WHO is determined as a standardized ICD-10 code. Assume that an ICD-10-code in the dataset is ${c_{10}}^t$, and an ICD-10 code in the mapping table is c_{10}. Then, an *SR* based on the longest common subsequence (*LCS*) [23], which examines the information changes in the dataset during the process using distance metrics is obtained as follows:

$$SR = 1 - \frac{|LCS(c_{10}, c_{11})|}{\max(n, m)}, \tag{1}$$

where the *LCS* distance between c_{10} and c_{11} is $n + m - 2 \mid LCS(c_{10}, c_{11}) \mid$, n is the number of the bits in an ICD-10 code, and m is the number of bits in a mapped ICD-10 code. If $SR = 0$, then the two codes are similar; therefore, code standardization is unnecessary. Otherwise, a large *SR* means a considerable difference between two codes. Thus, the *SR* in Equation (1) validates the examination of the information changes in code standardization.

Validation with the *UR* refers to the utilization of the Shannon entropy to examine the degree of uncertainty of mapping results from standardized ICD-10 to ICD-11. Generally, improper mapping often leads to the corruption of dataset validity. An ICD-10 code is associated with multiple ICD-11 codes, with their corresponding definitions. For example, an ICD-10 code "I25.1" in the mapping table is associated with two ICD-11 codes "BA80" and "BA8Z." The example prevents seamless transition from ICD-10-coded datasets to the ICD-11-based context. To evaluate the degree of uncertainty in such cases, we used *UR* in Equation (2) to illustrate the entropy changes in the process. Assume that an ICD-10 code c_{10} is ready to map a possible set of ICD-11 codes $\{c_{11}(i) \mid 1 \leq i \leq M\}$, where M is the number of ICD-11 code candidates, and a set of probabilities of the ICD-11 codes is $\{p_i \mid 1 \leq i \leq M\}$. According to the Shannon entropy, the *UR* of the $c_{10} \rightarrow c_{11}$ process between codes can be expressed as follows:

$$UR_{c_{10} \rightarrow c_{11}} = \sum_{1}^{M} p_i I(c_{10} \rightarrow c_{11i}) = \sum_{1}^{M} p_i \log \frac{1}{p_i}, \tag{2}$$

where $I(c_{10} \rightarrow c_{11})$ is the self-information of the $c_{10} \rightarrow c_{11}$ process, and the default log algorithm is $\log_2$. If $M = 1$, then $UR = 0$; thus, the ICD-10 code is precisely mapped to the correct ICD-11 code. The probability p_i in Equation (2) varies from the context of ICD-11. If all probabilities of ICD-11 codes are equal, then the *UR* in Equation (2) reaches a maximum entropy $\log_2 N$, where $\forall p = 1/N$.

The maximum *UR* in Equation (2) indicates that the mapping result reaches the greatest uncertainty [24], which refers to the most probable inaccurate transition on ICD-10-coded datasets, because the framework cannot determine an optimal choice based on equivalent probabilities. Therefore, some rules in differentiating each p_i are necessary. The *FC* in ICD-11 provides the rules by transforming the guidelines of ICD-11 coding into logical rules used in a coding system [25].

To increase the sensitivity of *UR* in evaluating uncertainty, we considered three cases, namely, Cases 1–3. In Case 1, the p_i in Equation (2) is assigned an average probability as in $1/M$. In Case 2, considering the text similarity of titles of ICD codes based on normalized Levenshtein metrics [26] provides additional information in determining the p^i in Equation (2). In Case 3, the p^i in Equation (2) is assigned randomly, assuming we do not consider the context factors of ICD-11. We denote the determination of probabilities in the three cases as a function *prob*(). Therefore, we have a multiple-code selection algorithm, as shown in Algorithm 1.

Algorithm 1. Multiple-code selection algorithm

Input: an ICD-10 entity c_{10}, a list of ICD-11 entities $\{c_{11}(i) \mid 1 \leq i \leq M\}$
Output: entropy of the process *UR*, optimal ICD-11 code c_{11}*
Let $pro \leftarrow \{p_i \mid 1 \leq i \leq M\}$ and $M \leftarrow$ the number of ICD-11 code candidates
for each $c_{11}(i)$ **in** $\{c_{11}\}$ **do**

$$p_i \overset{prob()}{\leftarrow} \begin{cases} 1/M, & \text{In } Case1 \\ NormalizedLevenshtein(c_{10}, c_{11}(i)), & \text{In } Case2 \\ Random(0,1), & \text{In } Case3 \end{cases}$$

end for
for each c_{11} **in** $\{c_{11}\}$ **do**
$p_i \leftarrow \frac{p_i}{\sum_1^M p_i}$
end for
return $UR = \sum_{i=1}^{M} p_i \log_2 \frac{1}{p_i}$, $c_{11}{}^* = c_{11}[i]$ where $s.t.\max(p_i)$

Finally, after a proper ICD-11 code $c_{11}{}^*$ to represent the original ICD-10 code c_{10} is determined using Algorithm 1, we have $IG(c_{10}, c_{11}{}^*)$ from code level, as follows:

$$IG(c_{10}|c_{11}{}^*) = I(c_{10}) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} I(c_{11}{}^*), \quad (3)$$

where $I(c_{10}) = UR$ is an entropy of an ICD-10 code to $c_{11}{}^*$, $I(c_{11}{}^*)$ is an entropy of the selected ICD-11 code $c_{11}{}^*$ in the context of ICD-11, S_v is the possible ICD-11 candidate in a mapping code set A, S is the total number of ICD candidates, and $|S_v|/|S|$ is the percentage of selected codes in the set. Thus, we have $IG(c_{10}, c_{11})$ bits of information regarding the code-level transition. Different *IG*s, such as IG_1, IG_2, and IG_3, for Cases 1–3, are obtained respectively. On the basis of Equations (1) to (3), a flowchart of the SV is developed, as shown in Figure 2.

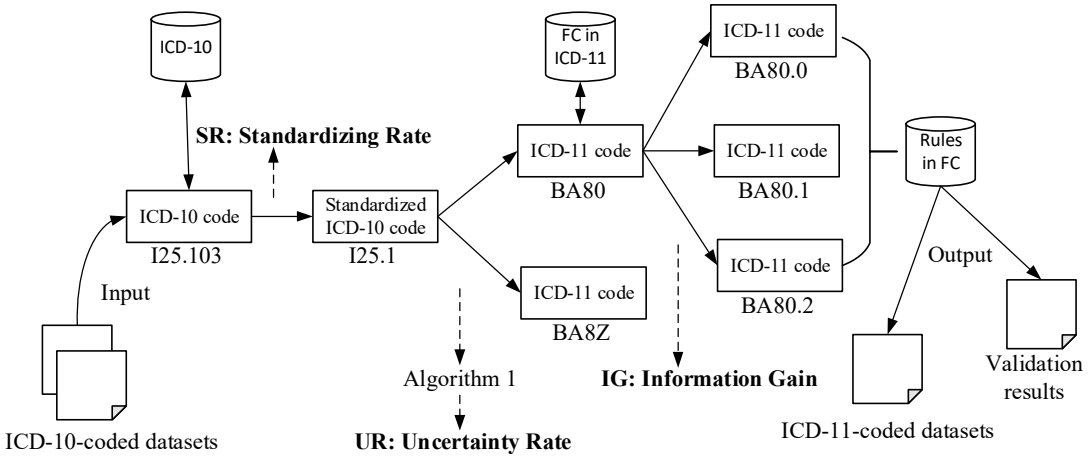

Figure 2. Flowchart of the Single-Code Validation (SV) in a mapping framework from ICD-10 to ICD-11.

In summary, we proposed the SV to validate the transition between ICD-10 and ICD-11 from code level. The *SR*, *UR*, and *IG* metrics were proposed to evaluate the changes of entropy during the code-to-code process.

2.2.2. Single-Disease Dataset Validation (SDV)

Based on the SV with three metrics, we proposed SDV to validate the records in a dataset focused on a single type of disease or a group of similar diseases. For example, a single-disease dataset includes patients who suffer from liver disorders. The goal of validating single-disease datasets is to examine whether the mapped ICD-10 codes were in a subset of ICD-11 disease-based domains. One of the

methods of validating the dataset with ICD codes was through arithmetic averaging of the *UR*s in the SV. Assume that H_c represents a *UR* obtained in the SV. Then,

$$H_{sd} = \frac{\sum_1^N H_c}{N}, \tag{4}$$

where N is the number of records in a single-disease dataset *sd*. However, the methods of evaluating the *sd* have drawbacks because the average entropy of ICD codes in a dataset cannot represent the entire distribution of validation results within a dataset. Another method to analyze *sd* is to evaluate the sample distribution of the entropy calculated in Equation (1). Suppose that H_{sd} is a random variable x. We use the cumulative distribution function (CDF), that is, $F_x(x) = P(X \leq x)$, to describe the probability distribution of the dataset. The CDF that illustrates percentage of samples provides an intuitive charting tool to validate the mapping results in a single-disease dataset.

We can also use a single metric to validate the effectiveness of the mapping process. By selecting a confidence interval e of *UR*, the distribution of entropy within a single-disease dataset can be depicted. Assume that an entropy is in a range of $e = [e_{\min}, e_{\max}]$, where e_{min} and e_{max} represent the minimum and maximum values of entropy, respectively. If a mapping result falls in e, then we represent the mapping accuracy for the dataset as follows:

$$accuracy = \frac{n_c}{n_c + n_e}, \tag{5}$$

where n_c is the number of the properly mapped record, and n_e is the number of incorrectly mapped records.

2.2.3. Multiple-Disease Dataset Validation (MDV)

An ICD-10-coded dataset with multiple diseases may have an influence on validation results. For example, ICD codes of chapters have different influences in the entropy during mapping. On the basis of the SDV, we proposed cluster-based MDV. Thus, we have the entropy of cluster i in a multiple-disease dataset as follows:

$$H(c_i) = -\sum_{j=1}^{|diseases|} p_{i,j} \log\left(p_{i,j}\right), \tag{6}$$

where $p_{i,j} = m_{i,j}/m_i$ is a probability that a record from cluster i belongs to class j, in which $m_{i,j}$ is the number of instances in cluster i with class, and m_i is the number of records in cluster i. Then, the average entropy of a clustering is obtained as follows:

$$\overline{H(c)} = \sum_1^K \frac{m_i}{m} H(c_i), \tag{7}$$

where m_i is the number of ICD codes in cluster i, and m is the number of all records in the dataset. In the MDV, the IG of the transition $IG(d,\{d_s\})$ can be expressed as

$$IG(d, \{c\}) = H(d) - \overline{H(c)}, \tag{8}$$

where $H(d)$ is the entropy of the multiple-disease dataset using the SDV.

In summary, on the basis of the SV and the SDV, the MDV allows the mapping framework to validate massive records of datasets with multiple diseases by considering the difference of multiple diseases in the context of ICD-11.

3. Results

3.1. SR during Standardization

We examined the *SR* on evaluating information changes when transforming the codes into standardized ones by analyzing 377,589 coded records in our test dataset. Figure 3 illustrates an empirical CDF of the *SR* in the dataset.

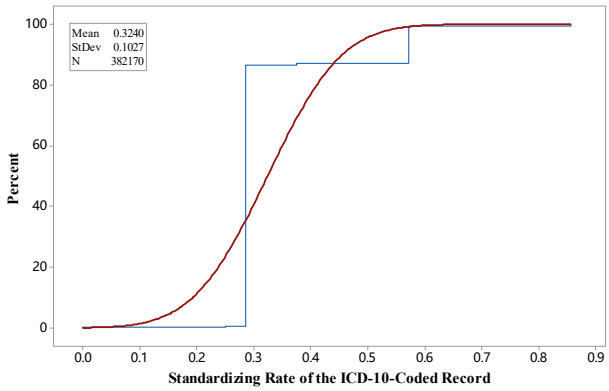

Figure 3. Empirical cumulative distribution of frequency of standardizing rate (*SR)* in an ICD-10-coded dataset.

Generally, large *SR*s refer to a considerable difference between an original code and a standardized one. Thus, when *SR* = 0, no difference between the two codes is observed. As shown in Figure 3, most of *SR*s fall into the range of [0.0, 0.6]. The result of *SR*s demonstrated that in ICD mapping, a considerable difference (of 60% of records) between the two codes existed, thereby affecting the efficiency of ICD mapping.

3.2. UR during Validation

On the basis of the standardized results in Figure 3, we implement the SV in the validation of the test dataset. We analyzed the experimental results of the *UR* to examine the aforementioned three cases in the mapping process. Figure 4 presents the Pareto chart of the entropy interval of Cases 1–3 in the SV.

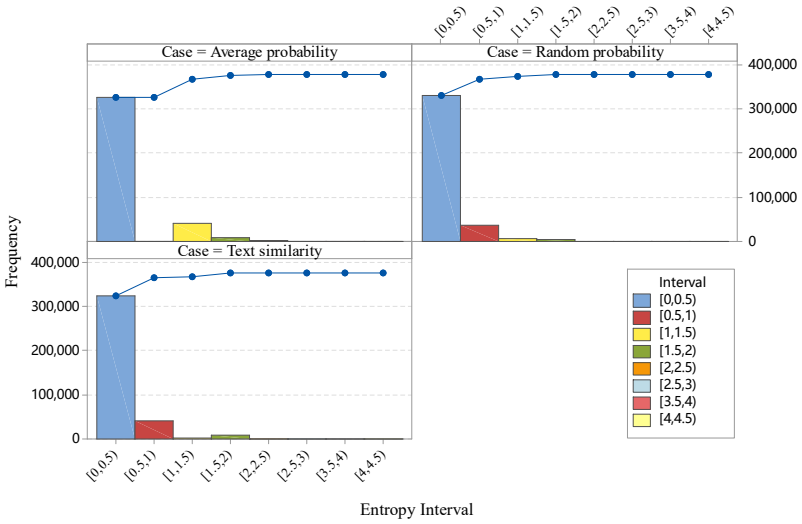

Figure 4. Pareto chart of the entropy interval of the three cases in SV.

The entropy-based *UR* in the SV provides a metric to examine the degree of uncertainty during the mapping process between two coding systems. Large *UR*s indicate considerable uncertainty of transition. As shown in Figure 4, three cases are considered. The difference of the cases in probability calculation of selecting ICD-11 codes was analyzed. As shown in Figure 4, *UR*s with 86.21% ICD-coded records of the dataset in all three cases are in the range of [0.0, 0.5], which indicates that the dataset achieves high *UR*s and can be mapped to ICD-11 codes precisely. However, 17.79% of the ICD-10-codes cannot be mapped to ICD-11 codes precisely. The comparison of the Pareto charts of the three cases shows that the *UR* in [0.0, 0.5] is constant, whereas the distribution of *UR*s in [0.5, $+\infty$] varies. Therefore, the ICD-codes with high *UR*s should be examined further for future use in an automated ICD mapping process.

3.3. IG during Mapping

Given that major changes exist during the revision of ICD-11, information changes of the mapped ICD-10 codes in the new context of ICD-11 must be validated. Based on the definitions of *IG*s in Equation (8), if an *IG* is less than zero, then the information of the mapped ICD-11 code contains more than that of the ICD-10 code. Table 1 provides the experimental results of *IG*s by analyzing the dataset. We divided the entire range of *IG*s gained in the process into 10 intervals of entropy and performed statistics on each interval.

Table 1. Frequency of different ranges of information gain (IG) during mapping.

Interval of Information Gain	Frequency	Percentage (%)	Number of ICD-10 Codes
$[-12, -10)$	86	0.00	2
$[-10, -8)$	0	0.00	/
$[-8, -6)$	0	0.00	/
$[-6, -4)$	63,256	0.17	107
$[-4, -2)$	101,378	0.27	523
$[-2, 0)$	50,695	0.13	264
$[0, 2)$	156,884	0.42	1405
$[2, 4)$	5277	0.01	92
$[4, 6)$	12	0.00	4
$[6, 8)$	1	0.00	1

For example, given 86 coded records with two distinctive ICD-10 codes in the dataset mapped to ICD-11 codes, the *IG*s fell in the range of $[-12, -10]$. The frequencies in Table 1 show a different distribution of the ICD codes in different intervals of *IG* during mapping. Approximately 57% of the records in the dataset tended to reduce uncertainty during the mapping from ICD-10 to ICD-11, whereas the other records provided further information to an ICD-11 coding system. The *IG*s of approximately 50% of the records were in $[0, 2]$, which demonstrated that the transition from ICD-10 to ICD-11 was likely to be in the same information amount.

Figure 5 shows an overview of the distribution of the number of ICD-10 codes in the dataset from different chapters, based on the results in Table 1. The number of ICD-10 codes changes in different chapters, over varying intervals of *IG*s. The results indicated that different multiple disease-based datasets may need to consider clustering as a first step to enable effective validation of the mapping process.

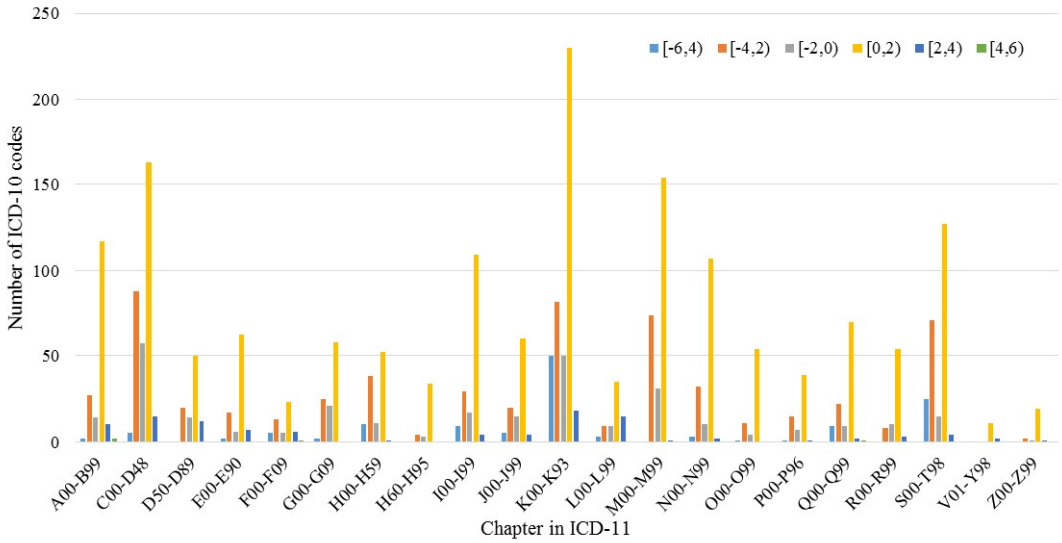

Figure 5. Changes of the number of ICD-10 codes in different chapters of ICD-10, over varying information-gain intervals.

4. Discussion

We proposed an ICD mapping framework and developed validation methods by leveraging existing tables from ICD-10 and ICD-11. Our method was a novel approach because we developed a mapping framework which uses existing tables of ICD-10 and ICD-11, to help utilize the existing ICD-10-coded datasets in adapting new features of ICD-11. We also proposed validation methods for evaluating information changes of codes in the mapping process during the transition, from the perspective of single-code, single-disease, and multiple-disease datasets.

As previously stated, our method aims to provide a new way to map existing ICD-10-coded datasets to the context of newly designed ICD-11 standards. In such mapping processes, we use the Shannon theory to monitor information changes of ICD codes from single-code, single-disease, and multiple-disease levels. We obtained several findings from the experimental results. First, ICD-10 codes from the datasets for the same diseases/disorders validated using *SR* may not be in tune due to different ICD-10 implementation in different hospitals and countries. The results in Figure 3 utilized a CDF chart to illustrate such a distribution. Second, during an automated mapping process, great complexity between two coding systems was observed, such as mapping from SNOMED CT to ICD-10-CM [27] or from ICD-9 to SNOMED CT [28]. *UR* was used to quantify such information changes during the transition from ICD-10 to ICD-11 codes. Then, we illustrated three cases of determining probability based on average probability, text similarity, and random probability integrated by Algorithm 1. A huge portion of the ICD-10 codes could seamlessly be mapped to ICD-11 codes, whereas the others with high uncertainty exist in the mapping process. The small portion of the codes in Figure 3 were the key to fulfill the proper transition between ICD-coded datasets. Ideally, the transition should be seamless. However, some codes in ICD-11 context change a lot. The key to a successful transition between the datasets is to ensure proper transition between the codes with great difference. Thus, a small portion of failed transition will cause the total failure of the use of an entire medical dataset. It shows the importance of the *SR* in evaluating the similarity between codes. Verifying the ICD-10 codes with high *UR*s can be useful to have a smooth transition. When an ICD-10 code is mapped to a new ICD-11 code, the new coding scheme in the context of ICD-11 provides more details of expert knowledge than that in ICD-10, which leads to inconsistency in the medical information between two coding systems. Moreover, the validation in single- and multiple-disease levels can be different in the evaluation of *IG* between codes from different ICD standards [29]. The findings from Table 1 and Figure 5 can help instruct future applications of ICD-11.

The release of ICD-11 in 2018 will revolutionize the development of innovation, technology, and application of medical informatics in the future [30]. Although ICD-11 is valuable to research on healthcare-related diseases, implementing the completely new ICD-11 standard in each member state of the WHO is difficult. ICD-9-CM coding alone may be insufficient to identify some specific diseases. For example, given that ICD-10 was first released more than 20 years ago, relatively more than 100 countries have reached the ICD-10 standards because the number of codes had increased from 13,000 in ICD-9 to 68,000 codes in ICD-10. In addition, doctors' workloads had increased after the adoption of ICD-10 because patients' diseases, diagnoses, and treatments must be recorded as accurately and precisely as possible. Meanwhile, medical institutions had to upgrade their healthcare information systems to adapt to the needs of ICD-10 coding. Substantial time and money were required to hire staff in the fields of medical research, information technology, and administration to complete the transition from ICD-9 to ICD-10. At present, ICD-11 contains approximately 269,280 codes. This number is greater than that in ICD-9 and ICD-10. It is believed that the difficulties and cost of promoting the use of ICD-11 at the beginning will also increase, but the patients and hospitals may benefit a lot through the improvement of the management level of medical informatics in the future. The revision of the existing coding systems should be automated, and redesigning the existing systems will affect the efficiency of ICD coders in the future. Therefore, utilizing and migrating existing ICD-10-coded data to a new context of ICD-11-based digital environments is likely to promote future use of historical medical data.

5. Conclusions

We proposed an ICD mapping framework for utilizing ICD-10-coded datasets to adapt to the new context of ICD-11. Three metrics, namely, *SR*, *UR*, and *IG*, to validate the information changes of ICD codes from the perspective of different levels in the framework were proposed. The *SR* is feasible to evaluate the information loss of ICD-10 codes, when the codes are standardized using a different method. Then, the results of calculating *UR* in the multiple-disease dataset indicated that the *UR*s of the datasets in the mapping process vary from different contexts in the ICD. We also examined the *IG* between ICD-10 codes and mapped ICD-11 codes. Our findings showed that the mapping process caused increasing uncertainty in selecting a proper candidate of ICD-11 codes related to an ICD-10 code, which should be considered when utilizing mapping and migration between two ICD coding systems in the future.

Author Contributions: Methodology, D.C.; Supervision, R.Z.; Writing–original draft, D.C.; Experimental design, X.Z.; Writing-review & editing, X.Z.

Funding: This work was partially supported by a key project of National Natural Science Foundation of China (Grant number 71532002) and a key project of Beijing Social Science Foundation Research Base (Grant number 18JDGLA017).

Acknowledgments: The authors thank the China Scholarship Council for their support of the scholarship.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. ICD-11 Home Page. Available online: https://icd.who.int (accessed on 16 July 2018).
2. Southern, D.A.; Hall, M.; White, D.E.; Romano, P.S.; Sundararajan, V.; Droesler, S.E.; Pincus, H.A.; Ghali, W.A. Opportunities and challenges for quality and safety applications in ICD-11: An international survey of users of coded health data. *Int. J. Qual. Health Care* **2016**, *28*, 129–135. [CrossRef] [PubMed]
3. Rodrigues, J.M.; Schulz, S.; Rector, A.; Spackman, K.; Millar, J.; Campbell, J.; Ustün, B.; Chute, C.G.; Solbrig, H.; Della Mea, V.; et al. ICD-11 and SNOMED CT Common Ontology: Circulatory system. *Stud. Health Technol. Inform.* **2014**, *205*, 1043–1047. [PubMed]
4. WHO Released New ICD-11. Available online: http://iogt.org/news/2018/06/20/who-released-new-icd11 (accessed on 15 July 2018).

5. Tudorache, T.; Nyulas, C.I.; Noy, N.F.; Musen, M.A. Using semantic web in ICD-11: Three years down the road. *Lect. Notes Comp. Sci.* **2018**, *8219*, 195–211.
6. Boerma, T.; Harrison, J.; Jakob, R.; Mathers, C.; Schmider, A.; Weber, S. Revising the ICD: Explaining the WHO approach. *Lancet* **2016**, *388*, 2476. [CrossRef]
7. Sofia, D.A.; Chris, F.; Shana, P.; Suarez-Almazor, M.E. Validation of ICD-9-CM codes for identification of acetaminophen-related emergency department visits in a large pediatric hospital. *BMC Health Serv. Res.* **2013**, *13*, 72.
8. Rey, G.; Bounebache, D.; Rondet, C. Causes of deaths data, linkages and big data perspectives. *J. Forensic Leg. Med.* **2016**, *57*, 37–40. [CrossRef] [PubMed]
9. Seare, J.; Yang, J.; Yu, S.; Zarotsky, V. Building a bridge: ICD-9-CM to ICD-10-CM mapping challenges and solutions. *Val. Health* **2014**, *17*, A187. [CrossRef]
10. Plznyak, V.; Reed, G.M.; Medina-Mora, M.E. Aligning the ICD-11 classification of disorders due to substance use with global service needs. *Epidemiol. Psychiatr. Sci.* **2017**, *27*, 212–218. [CrossRef] [PubMed]
11. ICD-11 Reference Guide. Available online: https://icd.who.int/browse11/content/refguide.ICD11_en/html/index.html (accessed on 16 July 2018).
12. Tu, S.W.; Bodenreider, O.; Çelik, C.; Chute, C.G.; Heard, S.; Jakob, R.; Jiang, G.; Kim, S.; Miller, E.; Musen, M.A.; et al. A content model for the ICD-11 revision. Available online: https://www.researchgate.net/publication/267792997_A_Content_Model_for_the_ICD-11_Revision (accessed on 2 October 2018).
13. Jiang, G.; Pathak, J.; Chute, C.G. Formalizing ICD coding rules using Formal Concept Analysis. *J. Biomed. Inform.* **2009**, *42*, 504–517. [CrossRef] [PubMed]
14. Fung, K.W.; Richesson, R.; Smerek, M. Preparing for the ICD-10-CM transition: Automated methods for translating ICD codes in clinical phenotype definitions. *eGEMs* **2016**, *4*, 1211. [CrossRef] [PubMed]
15. Brocco, S.; Vercellino, P.; Goldoni, C.A.; Alba, N.; Gatti, M.G.; Agostini, D.; Autelitano, M.; Califano, A.; Deriu, F.; Rigoni, G.; et al. "Bridge coding" ICD-9, ICD-10 and effects on mortality statistics. *Epidemiol. Prev.* **2010**, *34*, 109–119. [PubMed]
16. Shannon, C.E. A mathematical theory of communication. *Bell Sys. Techn. J.* **1948**, *27*, 379–423. [CrossRef]
17. Utter, G.H.; Cox, G.L.; Atolagbe, O.O.; Owens, P.L.; Romano, P.S. Conversion of the agency for healthcare research and quality's quality indicators from ICD-9-CM to ICD-10-CM/PCS: The process, results, and implications for users. *Health Serv. Res.* **2018**, *53*, 3704–3727. [CrossRef] [PubMed]
18. Startsev, N.; Dimov, P.; Grosche, B.; Tretyakov, F.; Schüz, J.; Akleyev, A. Methods for ensuring high quality of coding of cause of death. the mortality register to follow southern Urals populations exposed to radiation. *Methods Inf. Med.* **2015**, *54*, 359–363. [PubMed]
19. Simard, M.; Sirois, C.; Candas, B. Validation of the combined comorbidity index of Charlson and Elixhauser to predict 30-day mortality across ICD-9 and ICD-10. *Med. Care* **2018**, *56*, 441–447. [CrossRef] [PubMed]
20. Boyd, A.D.; Li, J.J.; Kenost, C.; Joese, B.; Yang, Y.M.; Kalagidis, O.A.; Zenku, I.; Saner, D.; Bahroos, N.; Lussier, Y.A. Metrics and tools for consistent cohort discovery and financial analyses post-transition to ICD-10-CM. *J. Am. Med. Inform. Assoc.* **2015**, *22*, 730–737. [CrossRef] [PubMed]
21. Quan, H.; Moskal, L.; Forster, A.J. International variation in the definition of "main condition" in ICD-coded health data. *Int. J. Qual. Health Care* **2014**, *26*, 511–515. [CrossRef] [PubMed]
22. Sundararajan, V.; Romano, P.S.; Quan, H.; Burnand, B.; Drösler, S.E.; Pincus, H.A.; Ghali, W.A. Capturing diagnosis-timing in ICD-coded hospital data: Recommendations from the WHO ICD-11 topic advisory group on quality and safety. *Int. J. Qual. Health Care* **2015**, *27*, 328–333. [CrossRef] [PubMed]
23. Chen, Y.; Lu, H.; Li, L. Automatic ICD-10 coding algorithm using an improved longest common subsequence based on semantic similarity. *PLoS One* **2017**, *12*, e0173410. [CrossRef] [PubMed]
24. Chako, S.J.; Danziger, R.; Boyd, A. Identifying clinically inaccurate conversions from ICD-9 to ICD-10 in cardiology clinical practice. *Circulation* **2014**, *130*, A11693.
25. Farkas, R.; Szarvas, G. Automatic construction of rule-based ICD-9-CM coding systems. *BMC Bioinformatics* **2008**, *9*, 1–9. [CrossRef] [PubMed]
26. Li, Y.; Liu, B. A normalized Levenshtein distance metric. *IEEE Trans. Pattern Ana. Mach. Intell.* **2007**, *29*, 1091–1095.
27. Cartagena, F.P.; Schaeffer, M.; Rifai, D.; Doroshenko, V.; Goldberg, H.S. Leveraging the NLM map from SNOMED CT to ICD-10-CM to facilitate adoption of ICD-10-CM. *J. Am. Med. Inform. Assoc.* **2015**, *22*, 659–670. [CrossRef] [PubMed]

28. Nadkarni, P.M.; Darer, J.A. Migrating existing clinical content from ICD-9 to SNOMED. *J. Am. Med. Inform. Assoc.* **2010**, *17*, 602–607. [CrossRef] [PubMed]
29. Khokhar, B.; Jette, N.; Metcalfe, A.; Cunningham, C.T.; Quan, H.; Kaplan, G.G.; Butalia, S.; Rabi, D. Systematic review of validated case definitions for diabetes in ICD-9-coded and ICD-10-coded data in adult populations. *BMJ Open* **2016**, *6*, e009952. [CrossRef] [PubMed]
30. Chui, K.; Alhalabi, W.; Pang, S.; Pablos, P.O.; Liu, R.W.; Zhao, M. Disease diagnosis in smart healthcare: innovation, technologies and applications. *Sustainability* **2017**, *9*, 1209. [CrossRef]