The Use of Analytic Hierarchy Process for Measuring the Complexity of Medical Diagnosis

Ofir Ben-Assuli\textsuperscript{1}, Nanda Kumar\textsuperscript{2}, Ofer Arazy\textsuperscript{3} and Itamar Shabtai\textsuperscript{4}

\textsuperscript{1} Faculty of Business Administration, Ono Academic College, Kiryat Ono, Israel.

\textsuperscript{2} Zicklin School of Business, Baruch College, NY, USA

\textsuperscript{3} Department of Information Systems, University of Haifa, Haifa, Israel

\textsuperscript{4} School of Economics and Management, College of Management, Rishon LeZion, Israel.
The Use of Analytic Hierarchy Process for Measuring the Complexity of Medical Diagnosis

Abstract

Diagnostic complexity is an important contextual factor affecting a variety of medical outcomes. Existing measurements of diagnosis complexity either rely on crude proxies or use fine-grained measures that employ indicators from proprietary data that are not readily available. Hence, the study of this important construct in fields such as medical informatics has been hampered by the difficulty of measuring diagnostic complexity.

This paper presents a novel approach for conceptualizing and operationalizing diagnostic task complexity as a multi-dimensional construct, which employs the readily available ICD codes from medical encounters in hospitals and uses Analytic Hierarchical Process (AHP) methodology. We demonstrate the reliability of the proposed approach and show that despite using a relatively simple procedure, it is able to predict re-admission rates just as well as (or even better) than some of the sophisticated measures that have been used in recent studies (namely, the LaCE score index).

Keywords: Medical Informatics; Task Complexity; Diagnosis Complexity; Analytic Hierarchy Process (AHP); International Classification of Diseases (ICD).
The Use of Analytic Hierarchy Process for Measuring the Complexity of Medical Diagnosis

1. Introduction

Complexity of medical diagnosis is an important contextual factor that can impact a variety of outcomes in the medical domain, including quality of care during the medical encounter, as well as hospitalization outcomes such as readmission rates (1). Diagnosis task complexity is a crucial moderator variable that can explain the link between physicians’ use of information systems and patient outcomes (2-6). Notwithstanding the importance of this construct, complexity in general has proven to be an elusive construct to measure across various settings (7).

In particular, conceptualization and measurement of diagnostic complexity in medical settings is especially challenging due to the unpredictable nature of doctor-patient encounters and the multitude of contextual factors that make an encounter more or less complex. This difficulty is further exacerbated by the sheer volume of data captured about a clinical encounter between a patient and a doctor (8). Existing approaches to measuring diagnosis complexity either rely on crude proxies such as the number of diagnoses and risk of complications, or alternatively employ techniques that use a wide array of indicators (as in the case of the severity index, Elixhauser and Charlson/Deyo Methods of Comorbidity Measurement (9, 10), or the LaCE score index (11, 12)). The methods that use a large number of clinical indicators (anywhere from 15 to 40 indicators) are not always readily available to practitioners and researchers. Critically, these more elaborate measures are still proxies for complexity as they all measure constructs specifically meaningful for clinical use (such as severity of illness, co-morbidity index among others).

Hence, the primary objective of this article is to suggest a middle-ground approach for operationalizing and measuring medical diagnosis complexity. We develop this approach in a manner that captures the multidimensionality of the diagnosis task complexity construct, yet requires a simpler procedure when
compared to the complicated procedures that are employed to produce metrics such as those mentioned above.

Moreover, our approach relies on data that is commonly available, namely the ICD diagnosis codes, a system of coding patient-physician encounters to record diagnoses. The use of ICD codes has advantages over other indicators captured in hospitals’ computerized systems. Past research has shown that accuracy of hospitals’ data for previous comorbidity codes tend to be very poor for a variety of reasons (13). The factors contributing to the unreliability of these co-morbidity data include poor IT infrastructure leading to lack of information sharing, financial pressures (for better remuneration by insurance companies) leading to dysfunctional coding of co-morbidities among others (10, 14, 15). This difficulty is further exacerbated by the absence of adequate electronic data exchange between the Electronic Medical Record (EMR) and other clinical departments (lab, radiology, and referral systems) (16). In contrast, ICD codes are recorded at the time of patient-physician encounter, and thus are less susceptible to these concerns.

ICD codes have been employed to measure an array of clinical constructs in the medical domain, including task complexity. For example, the Elixhauser and Charlson methods (17) uses ICD diagnosis codes for comorbid conditions (presence of other diseases in addition to the primary disease the patient came in to get treated) that are present before the patient encounter takes place. It is important to note that many of the methods that employ ICD codes are primarily interested in broader clinical outcomes for the patients or hospitals, and are not intended to produce a metric for the particular medical encounter. For example, the primary objective of the Charlson method is to predict mortality of the patient in a specific timespan. Few approaches that attempted to characterize patient-physician encounter have relied on an array of indicators beyond ICD codes, such as diagnosis and physiological indicators (e.g. the severity index) or medications data from pharmacy systems (e.g. the chronic disease score).

Using a procedure that employs the AHP methodology, we are able to “translate” ICD codes into measures for several theoretical dimensions of complexity (namely, diagnostic difficulty, the amount of work done, and risk of complications). The proposed procedure is relatively simple and requires few
resources. Given that task complexity is a key construct in the medical informatics field (18, 19), offering a straightforward approach to measuring the construct along various dimensions has the potential to advance research in the area, for example by enabling future research to explore the linkage between task complexity and a variety of outcomes (e.g. quality of care or EHR adoption).

The second objective of this paper is to present the use of a research methodology - AHP – to informatics scholars, using the conceptualization, operationalization and measurement of diagnostic complexity as an example. AHP is a multi-criteria decision making (MCDM) method that is very flexible and easy to use. It structures a problem hierarchically over multiple levels of criteria that are often conflicting and subjective (20). This AHP procedure begins with problem formulation/modeling. For example, this stage may entail examining complexity in terms of multiple levels of subjective and objective criteria (for instance Ben-Assuli (21)). The formulation stage is then followed by operationalization, measurement and interpretation. AHP can be used in health technology assessment to give a quantitative dimension to physicians’ decisions or patients’ preferences for treatment endpoints. Although originally developed to support decision making, AHP may also play a role in identifying or prioritizing the net benefit of health interventions. AHP is also useful for policy makers in combining multiple clinical outcomes of healthcare interventions in an overall health economic evaluation (22, 23). AHP may be particularly relevant in cases where different outcome measures lead to conflicting results about the best alternative to reimburse. As we later show, AHP is a very good candidate for assessing the complexity of medical diagnostic decisions. Furthermore, the AHP procedure can be generalized to other constructs in the informatics domain, and our hope is that the description here will offer scholars in the field with a quantitative approach for assessing a host of phenomena.

In what follows, we first highlight the importance of the diagnostic complexity construct and review empirical studies of patient outcomes to stress the challenges associated with the measurement of this multifaceted construct. We then introduce the AHP method and show how it can be employed to derive from commonly available ICD codes a score for the diagnostic complexity of doctor-patient encounters.
along several key dimensions (diagnostic difficulty, the amount of work done, and risk of complications). We present analyses demonstrating the internal consistency of our AHP-based complexity measures, and show that when multiple independent assessors estimate complexity the inter-rater agreement levels resulting from the AHP procedure are superior to those obtained through a standard Likert-scale assessment. To illustrate the effectiveness of our complexity scales we demonstrate that they could serve to predict important clinical outcomes. In particular, we compare our AHP-based complexity measures to the LaCE score (11, 12) that was designed to predict the risk of readmission (L-length of stay, a- age of the patients, C- Charlson comorbidity score index and E- number of previous ED encounters in the previous half year), and show that our scale - although simple to produce and require only ICD codes – performs just as well as the complex LaCE index. In order to promote future research in the area, we make public our data regarding the complexity scores – along the various dimensions – that associated with the ICD codes used in this study (high-level 3-digit code categories for some of the most common codes).

In summary, our main goal is to demonstrate a novel approach for conceptualizing and operationalizing diagnostic task complexity as a multi-dimensional construct, as well as to show that this AHP-based approach can be used to tackle similar multi-criterion problems in healthcare.

1.1 Background and Significance

The task of medical diagnosis is complex and is prone to complications (24). According to National Academies of Sciences and Medicine (25), diagnostic errors have the potential to severely harm patients, and thus it is essential that the diagnosis process be improved. As Angus et al. (26) explain, complex diagnosis tasks are often associated with clinical side effects or complications, thus placing a burden on the healthcare budget.

1.1.1 Task Complexity

Historically, task complexity has been conceptualized in a multitude of ways that typically consist of dimensions such as amount of work to be done, degree of difficulty, degree of task structure and novelty
of work (27). Campbell (7) argued that uncertainty and degree of task structure play an important role in making a task more or less complex. He made the case that when conceptualizing complexity, attention needs to be paid to factors such as the presence of multiple paths to the desired end-state, the presence of multiple desired end-states, the presence of conflicting interdependence and the presence of uncertainty (or probabilistic linkages).

Consistent with work done in other disciplines, there is a broad consensus in medical research that complexity is a multi-dimensional construct. However, not enough measures of complexity have been established in the clinical domain (28-30). Typically, clinical measures have tended to emphasize the number of chronic conditions, medications and even the use of medical resources/costs as a proxy for complexity (29, 30). Further, commonly used proxies are not always accurate (31, 32). Kimberlin and Winterstein (31) argue that in medical research there is a need to better align theoretical concepts (for example, complexity) with proxy measures in appropriate contexts so that they are reliable and valid.

1.1.2 Revisiting the AHP Method

AHP is a multi-criteria method for organizing and analyzing complex decisions (20, 33-35). The AHP method, developed by Saaty (20), is aimed at solving problems with a large number of variables and is particularly useful when dealing with a combination of quantitative and subjective elements. The AHP model takes into account a judgment rather than an objective measurement, and assigns a meaning to the priorities made by the decision maker. The AHP procedure serves to structure the decision problem into a hierarchy reflecting the underlying values, goals and objectives of the decision makers (36). While AHP has been used in a handful of studies in the clinical context in the healthcare domain (34), to the best of our knowledge it has not been employed in the medical informatics field at the intersection of healthcare and technology, despite its potential.

AHP constitutes a measurement and comparison tool that determines the substitutability between pairs of criteria which relate to a decision-making process. Specifically, when using AHP, the decision maker
compares one pair of factors at a time, instead of scoring each factor directly (e.g. rating the element on a Likert-scale). Without a rigorous method such as AHP, the decision outcome is prone to biases (37).

Multiple comparisons between pairs of factors can lead to inconsistencies (38), such as the primacy effect, the recency effect, and the bandwagon effect (where people express their opinions based on what they expect other people think). These inconsistencies within AHP are controlled for and measured by an inconsistency coefficient. Lane and Verdini (39) recommended taking a stringent view of inconsistencies when the inconsistency coefficient exceeds 10%, which is the standard applied in the case study here.

1.1.3 Diagnostic Complexity: Development of Criteria

The first step in the AHP procedure entails problem formulation. In our context, we begin with a conceptualization of diagnosis complexity by decomposing the construct into different factors (as outlined below). We then show how our conceptualization can be operationalized with AHP by linking complexity to ICD codes.

One particularly relevant way of conceptualizing complexity in the clinical setting is to look to practice to see how providers are compensated for the provision of medical care. A widely used report from the Center for Medicare and Medicaid Services (40) specifically operationalizes the complexity of medical decision-making based on: a) the number of diagnoses or management options; b) amount and/or complexity of data to be reviewed; and c) risk of complications and/or morbidity. This categorization is in line with extent theoretical frameworks (41). The report uses these factors to derive an aggregate diagnostic complexity score along a continuum spanning four levels: straightforward, low-complexity, moderate-complexity and high complexity (see Table 1). This complexity scale is widely used in hospitals throughout North America to reimburse medical providers by the government for the care rendered during a medical encounter.
Table 1: Task Complexity Criteria (from CMS (40))

<table>
<thead>
<tr>
<th>Number of Diagnoses or Management options</th>
<th>Amount and/or Complexity of Data to be Reviewed</th>
<th>Risk of Complications and/or morbidity or mortality</th>
<th>Type of Decision Making</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimal</td>
<td>Minimal or None</td>
<td>Minimal</td>
<td>Straightforward</td>
</tr>
<tr>
<td>Limited</td>
<td>Limited</td>
<td>Low</td>
<td>Low Complexity</td>
</tr>
<tr>
<td>Multiple</td>
<td>Multiple</td>
<td>Moderate</td>
<td>Moderate Complexity</td>
</tr>
<tr>
<td>Extensive</td>
<td>Extensive</td>
<td>High</td>
<td>High Complexity</td>
</tr>
</tbody>
</table>

Synthesizing the complexity indicators fleshed out by Gill and Hicks (27)) together with the guidelines provided by CMS in evaluating clinical complexity, our operationalization of medical diagnosis complexity focuses on the following three dimensions as the top-level criteria:

(i) Diagnosis difficulty (e.g. a diagnosis can be harder to arrive at if the patient presents symptoms that suggest multiple candidates for diagnosis or if the patient presents symptoms that are contradictory);

(ii) The amount of work done in diagnosing, and

(iii) Risk of complications or co-morbidity (reflecting the importance of the decision).

While in the case of our operationalization of diagnosis complexity, this top-level conceptualization provided sufficient granularity, it should be noted that the AHP procedure allows for a further breakdown of the hierarchical structure, possibly descending multiple levels.

1.1.4 ICD Codes and Diagnosis Complexity

We derived our diagnosis complexity metrics along the three dimensions mentioned above by turning to a commonly available, well-structured, standardized categorization schema (ICD), which is used to code and record medical diagnoses during doctor-patient encounters. The World Health Organization, the National Center for Health Statistics (NCHS) and the CMS have developed the "ICD" coding system, which has gone through a series of revisions and updates (42, 43). The ICD coding system was designed
to promote international comparability in the collection, processing, classification, and presentation of mortality statistics. It assigns codes to both diagnoses and procedures associated with hospital utilization in the United States (and around the world) and is the dominant standard for billing and morbidity indexing in the United States. This reporting responsibility makes the assignment of ICD codes a necessity for most or all healthcare organizations, and the ICD codes are commonly used in EHR as a means of recording the nature of the medical encounter (note that the ICD diagnoses coding scheme is most relevant to our investigation).

The availability of coded data has led to the use of ICD codes for research involving clinical records, where typically these codes are used to identify the incidence of a specific condition. Frequently, the data derived from the code are used to compute simple statistics – for example the prevalence of specific conditions such as cancer types or sepsis in order to utilize them in clinical research (44-46). For example, Calle, Rodriguez (45) investigated the putative role of obesity in increasing the risk of death in different types of cancers, where the latter was computed using ICD codes. The primary limitation of this approach is that it does not allow for a direct mapping between ICD codes and constructs such as diagnosis task complexity. An alternative methodology aims at developing measures for more general constructs by utilizing ICD codes. For example, scoring systems have been developed to evaluate the severity of a patient’s injuries using ICD codes (47). However, this approach requires additional granular details to map a particular ICD code to a construct category (e.g. high severity). Because of the extensive data collection and manual effort required in developing such a scheme, data owners often choose not to make it freely available, thus limiting the potential for research in this area.

Here we advocate a middle ground, where we map ICD codes to a theoretically grounded construct – in our particular case, diagnosis complexity – by relying on ICD codes alone. One immediate outcome of this straightforward approach for mapping of codes to constructs is that the mapping scheme is made publicly available. Our ICD code-construct mapping employs a rigorous and relatively underused methodology known as AHP.
2. Method and Results

This section describes the various steps in applying the AHP method including: problem formulation, articulation of decision hierarchy, construction of pairwise comparison matrices, consistency validation, and the computation of AHP score. In addition, we demonstrate the effectiveness of the AHP scales and compare their consistency against those of standard Likert scale ratings.

2.1. Using AHP to Derive Diagnosis Complexity Score: A Brief Demonstration of the AHP Method,

To arrive at a diagnostic complexity score for patient encounters, we apply the following steps (for a detailed description of the AHP procedure see Appendixes A and B):

2.1.1. Step 1: Formulate the Problem

In this step, we identified the problem as the need to evaluate the diagnostic complexity of patient encounters so that we can use it in empirical studies to evaluate a panoply of issues such as EHR system use, and quality of care. This step was discussed in more detail in Section 1.1 above.

2.1.2. Step 2: Conceptualize the Problem by Structuring the Decision Hierarchy

Next, to derive a conceptualization of task complexity we consulted a number of resources including the vast literature on task complexity and studies on diagnostic complexity in medicine, as well as practical guidelines. The result of this process was a three-dimensional conceptualization of diagnosis task complexity made up of diagnostic difficulty, the amount of work done, and risk of complications. We further evaluated the face validity of this conceptualization by interviewing experts in the area (four physicians in different specializations).

2.1.3. Step 3: Construct Sets of Pairwise Comparison Matrices

In this step, we relied on the ICD-9 codes associated with each patient record to evaluate the complexity of each diagnosis. We employed a large database of emergency department (ED) referrals. Although the EDs experience enormous pressure and constraints on manpower resources, there are positive findings for improved financial outcomes when EHR were implemented in EDs (48). The log-files were made up of
data collected from seven main hospitals, all owned by the primary Health Maintenance Organization (HMO) in Israel, over a period of four years. The EDs in these hospitals use an EHR over a Health Information Exchange (HIE) network platform to share medical information from points of care operated by their affiliated health suppliers. We worked with the ten most common codes in our dataset, as listed in Table 2. The mapping of ICD-9 codes onto these complexity dimensions involved a labor-intensive manual procedure employing the AHP technique (which yielded a final complexity score for each ICD code group).

<table>
<thead>
<tr>
<th>ICD9 Code Group</th>
<th>Description</th>
<th># records</th>
</tr>
</thead>
<tbody>
<tr>
<td>920-924</td>
<td>Contusion with Intact Skin Surface</td>
<td>160,487</td>
</tr>
<tr>
<td>840-848</td>
<td>Sprains and Strains of Joints and Adjacent Muscles</td>
<td>63,250</td>
</tr>
<tr>
<td>360-379</td>
<td>Disorders of the Eye and Adnexa</td>
<td>58,236</td>
</tr>
<tr>
<td>810-819</td>
<td>Fracture of Upper Limb</td>
<td>57,928</td>
</tr>
<tr>
<td>380-389</td>
<td>Diseases of the Ear and Mastoid Process</td>
<td>42,986</td>
</tr>
<tr>
<td>958-959</td>
<td>Certain Traumatic Complications and Unspecified Injuries</td>
<td>38,142</td>
</tr>
<tr>
<td>720-724</td>
<td>Dorsopathies</td>
<td>33,896</td>
</tr>
<tr>
<td>590-599</td>
<td>Other Diseases of Urinary System</td>
<td>31,855</td>
</tr>
<tr>
<td>880-887</td>
<td>Open Wound of Upper Limb</td>
<td>31,674</td>
</tr>
<tr>
<td>555-558</td>
<td>Non-infective Enteritis and Colitis</td>
<td>30,654</td>
</tr>
</tbody>
</table>

We then recruited experts to help convert the ICD codes into a diagnosis complexity scale. We asked four senior physicians to participate in the study (all with extensive ED experience in the hospitals included in our sample). Each physician was first informed about the aim of the study. Then to generate an aggregate complexity construct, they were provided with definitions of the three complexity dimensions and were asked to directly compare each of these dimensions in terms of their importance, as shown in the figures below (Figure 1A and Figure 1B). The physicians were asked use a computer-controlled sliding scale to indicate the relative importance of each dimension (on a scale of 9).
The physicians were then handed ten cards describing each of the ICD-9 code groups (for each group, its name and the list of specific codes it contained). For purposes of illustration, see Figure 2 below for illustration of one ICD-9 code group.

**Figure 2. An Example of a Card (for “Fractures of Lower Limb; ICD code group 820-829) Used in the AHP Pair-Wise Comparisons of ICD Code Groups.**

- 820 Fracture of neck of femur
- 821 Fracture of other and unspecified parts of femur
- 822 Fracture of patella
- 823 Fracture of tibia and fibula
- 824 Fracture of ankle
- 825 Fracture of one or more tarsal and metatarsal bones
- 826 Fracture of one or more phalanges of foot
- 827 Other multiple and ill-defined fractures of lower limb
- 828 Multiple fractures involving both lower limbs lower with upper limb and lower limb(s) with rib(s) and sternum
- 829 Fracture of unspecified bones

The physicians were then asked to compare each combination of codes in terms of the three complexity dimensions and record their comparisons. This resulted in three 10x10 matrices. The example below
reproduces one such matrix that resulted from a comparison of the 10 ICD codes to each other in terms of one of the complexity factors (e.g. diagnosis difficulty). The ICD codes are listed in the top row in Figure 3. The physician made 45 direct comparisons for this one matrix).

Figure 3. The Resulting AHP Matrix For the Ten ICD Codes (for a single assessor).

For example, the value of 3.0 in the (1x4) element of the matrix indicates that the participant thought that ICD code 840-848 was 3 times as important in determining diagnosis difficulty as ICD code 920-924. Given this entry, the software automatically computed the value for the transpose element (the relative importance of 920-924 compared to 840-848). The AHP method computes this by performing the simple calculation of $1/x$; hence, 3 results in 0.33 for its transpose; 4 results in 0.25 that is rounded off to 0.3 for this illustration.

2.1.4. Step 4: Check the Ratings in the Pairwise Comparison Matrices for Consistency

AHP then organizes the pairwise comparisons in matrix form. A weighted average can eventually be used to compute a final weight (complexity score) for each ICD code. But before deriving this score, the ratings need to be tested for two types of consistency – an individual consistency measure that examines whether ratings by a single individual are internally consistent and a group consensus measure that examines ratings between individuals to arrive at a consensus.

At this stage, we transferred the pairwise matrices from the ‘Expert Choice’ software (http://expertchoice.com/) we used to perform the comparison to MS Excel. AHP produces an internal
consistency metric by identifying when a subject’s pairwise comparisons are not entirely in line with one another. We checked the consistency of all four experts’ answers on each dimension and found that their responses more than adequately met the criteria for an absence of inconsistencies in judgment (49).

We used Excel after adjusting the values to solve minor inconsistency problems using Bunruamkaew (50). This method, involves rating each decision alternative and assigning the weights for each criterion by developing a pair-wise comparison matrix for each criterion, normalizing the resulting matrix, averaging the values to get the corresponding rating and then calculating and checking the consistency ratio then calculating the weighted average rating for each decision alternative (50-52). See Appendix A for a more detailed information on these steps.

The physicians who mapped the ICD-9 codes to the complexity constructs all met the accepted tolerance level for inconsistencies in AHP (Diagnosis Difficulty: 5.0-9.6%; Amount of Work: 5.1-9.8%; and Risk of Complications: 2.3-8.7%), thus providing us with reliable scores for diagnosis complexity. We then analyzed the consensus between the four different assessors to find the following consensus levels (calculated by AHP): Difficulty: 66.9%; Amount of Work: 71.7%; Risk: 67.5%. Based on the accepted interpretation for inter-agreement metrics (such as Kappa, or ICC), a consensus of 0.6-0.8 is interpreted as “High Agreement” (53, 54).

2.1.5. Step 5: Use these Pairwise Comparison Matrices to Compute the Final Score

Once the measures for each criterion had been calculated independently, AHP can aggregate these scores into the higher-level complexity construct. We used the Brunelli, Canal (49) procedure to aggregate the complexity dimensions (amount of work done, diagnostic difficulty, and risk of complications) into a single aggregate construct (55).

We then calculated the geometric prioritization row geometric mean method (RGGM) for each participant based on the literature (55) (see Appendix B). RGGM is a prioritization method that computes the final priority order for alternatives. We computed the weighted scores using the pairwise comparisons of all respondents across the ICD-9 diagnoses for each dimension using the process in Xu (56) (Figure 4). The
consolidated calculation is based on the geometric mean of the participants as shown below (n- The number of criteria; Xi- Decision values per criterion and Wi- Weighted matrix per criterion).

**Figure 4: The Weighted Calculation of All Assessors.**

\[
\bar{x} = \left( \prod_{i=1}^{n} x_i^{w_i} \right)^{1/\sum_{i=1}^{n} w_i} = \exp \left( \frac{\sum_{i=1}^{n} w_i \ln x_i}{\sum_{i=1}^{n} w_i} \right)
\]

We finally arrived at the weights by calculating the geometric prioritization (RGGM) of the different diagnoses for each dimension according to all respondents, including normalization according to Dong, Zhang (55). Figure 5 below shows the final weights (see Appendix B for the calculation steps).

**Figure 5: RGGM Output for All Diagnoses According to All Respondents.**

After verifying the level of consensus (see Appendix B) across all physicians, we calculated the AHP complexity score for each ICD-9 code as the average of the four assessors. To complete the analysis, each patient record was associated with a diagnosis complexity score based on the AHP score for the ICD-9 code in that patient's record. This complexity score can now be used to hypothesize the relationship between diagnostic complexity and a host of critical constructs in the health information technology.
(HIT) field (EHR use, task complexity as well as other relevant antecedents and other outcome variables such as quality of care).

2.1.6. Demonstrating the effectiveness of our AHP-based complexity scales

A key conjecture to our study is that our AHP-based scales of diagnosis task complexity provide a middle ground: on one hand they rely on few data sources (namely, the commonly available ICD codes) and on the other hand they are effective in the sense that they can replace the more sophisticated measures that are used in the literature. Thus, we computed from our data the LaCE score and compared it against the AHP-based measure in terms of its ability to predict a common outcome measure: readmission rates. The LaCE index (11, 12) has been developed to identify patients at high risk of hospital readmission, so they can be targeted for interventions aimed at reducing the rate of readmission (57). Computing the LaCE score demands additional effort for collecting more data (as explained in the Introduction section), whereas all the information needed for our AHP-based complexity index (namely, ICD codes) are publicly available. Our analysis shows that the correlation between LaCE and readmission rates were 0.008 (this figure is in line with findings from other studies that have found weak impact of LaCE on readmissions (58)). In comparison, the correlation for our AHP-based complexity index was substantially higher 0.017 (both correlations, although quite weak, were statistically significant, P-value < 0.001). This analysis demonstrates the usefulness of our approach for developing a task complexity index.

2.1.7. Analysis of inter-rater agreement: comparison AHP against Likert scale ratings

The analysis presented above demonstrate that for each of our study’s assessors independently, AHP scores (for the various complexity dimensions) were internally consistent. The additional analysis presented in this sub-section is intended to test the inter-rater agreement (or consistency) between the assessors. As a baseline, we asked our assessors to rate the complexity (along the various dimensions) of each of the ICD codes using a 5-point Likert-scale. We then compared inter-rater agreement levels between AHP and Likert scales for the same set of assessors. As can be seen in Table 3, for all three
complexity dimensions, AHP-based scales yielded substantially higher agreement levels when compared to the Likert scale, in terms of both Cronbach Alpha and itra-class correlation coefficients (ICC).

**Table 3. Inter-rater agreement comparison AHP vs. Likert over all three Complexity Dimensions**

<table>
<thead>
<tr>
<th>Construct</th>
<th>Method</th>
<th>Cronbach Alpha</th>
<th>ICC Single Measures</th>
<th>Significance</th>
<th>ICC Average Measures</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amount of Work</td>
<td>Likert</td>
<td>0.682</td>
<td>0.326</td>
<td>0.013</td>
<td>0.659</td>
<td>0.013</td>
</tr>
<tr>
<td></td>
<td>AHP</td>
<td>0.848</td>
<td>0.612</td>
<td>0.000</td>
<td>0.863</td>
<td>0.000</td>
</tr>
<tr>
<td>Difficulty</td>
<td>Likert</td>
<td>0.733</td>
<td>0.488</td>
<td>0.006</td>
<td>0.741</td>
<td>0.006</td>
</tr>
<tr>
<td></td>
<td>AHP</td>
<td>0.791</td>
<td>0.519</td>
<td>0.000</td>
<td>0.812</td>
<td>0.000</td>
</tr>
<tr>
<td>Risk of Complications</td>
<td>Likert</td>
<td>0.651</td>
<td>0.325</td>
<td>0.013</td>
<td>0.658</td>
<td>0.013</td>
</tr>
<tr>
<td></td>
<td>AHP</td>
<td>0.795</td>
<td>0.524</td>
<td>0.000</td>
<td>0.815</td>
<td>0.000</td>
</tr>
</tbody>
</table>

### 3. Discussion

Task complexity has always been a critical antecedent in explaining task performance (7, 27). With respect to clinical encounters between physicians and patients, the diagnostic complexity of that encounter should be a critical variable in explaining clinical outcomes. However, task complexity in general and diagnostic complexity in particular have always been difficult constructs to operationalize. Existing approaches to measuring diagnosis complexity either rely on very crude proxies, or alternatively employ an extensive set of indicators from proprietary data that are not readily available to practitioners and researchers and are still indirect proxies to measure complexity. As a consequence of the difficulty of measuring diagnostic complexity, the application of this important construct has been limited.

We proposed a promising middle ground approach to conceptualizing and measuring diagnosis complexity, which captures the various facets of this construct, yet requires relatively little manual effort and relies on readily accessible data. We reviewed research on diagnosis complexity as well as the ways in which expert physicians characterize these clinical encounters. Based on this exploration, we conceptualized complexity as a multi-dimensional construct comprised of diagnosis difficulty, the amount of work done in diagnosing and risk of complications/co-morbidity. The importance of our work lies in linking this conceptualization of complexity to ICD codes. We accomplish this by using an under-utilized
research method – AHP. We demonstrate that our AHP-based scales of diagnosis complexity is a useful predictor of readmission rates and could effectively be used to replace commonly used measurements of the clinical process, such as LaCE score index. By making our complexity scores for the most common ICD codes publicly available, we hope to facilitate future research that is interested in understanding the antecedents and consequences of diagnosis task complexity.

We provided a short demonstration of how to use the AHP method with the hope that practitioners will come to realize the potential of this approach, potentially applying it in a variety of medical domain to solve a range of problems that can be structured hierarchically. We used the AHP method to assign complexity scores for ten high-level ICD codes. Since AHP requires experts to provide comparisons among the different options (i.e. the ICD categories), the number of comparisons increase exponentially with the number of options. Thus, the usefulness of this approach is limited by the amount of manual labor. In our case, we employed this method to assign scores for the most popular ICD codes, which accounted for a significant portion of the records in our data. Applying AHP to problems that include a large number of options (and require many comparisons) would require the investment of substantial effort.

In conclusion, despite the importance of the diagnostic complexity construct, to date scholars and practitioners alike struggled with its measurement. Existing measurements of diagnosis complexity either rely on crude proxies or alternatively use fine-grained measures that employ indicators from difficult-to-come-by proprietary data. This study proposed a relatively simple approach for estimating diagnosis complexity by relying on readily available ICD codes, and a process based on the AHP method. The findings from our study – complexity scores along three important dimensions for a set of very common ICD codes – could be used in future research that attempts to link diagnosis complexity to other constructs.

Analytic Hierarchical Process (AHP) is a rigorous multi-criteria decision making method that works by structuring a problem hierarchically over multiple layers of criteria, making it possible to compare criteria
that can often appear subjective (for example, diagnosis difficulty and risk of complications). Although AHP is not a new approach and has been applied before in the medical domain (e.g. to make decisions on patient participation, diagnosis, technology evaluation and healthcare policy (34, 59)) and beyond (e.g. to operationalize project complexity where project complexity (60, 61)), its application has been rather limited. Medical research is challenged by the difficulty of assessing a variety of subjective phenomena. By demonstrating the value of the AHP method and providing a guide on how to use it, we hope to open the door for future research would apply AHP to the measurement of hierarchically-structured constructs in the medical domain.

4. Bibliography

10. van Walraven C, Austin PC, Jennings A, Quan H, Forster AJ. A modification of the Elixhauser comorbidity measures into a point system for hospital death using administrative data. Medical care. 2009:626-33.


42. Aalseth PT. Medical Coding: Jones & Bartlett Publishers; 2014.


50. Bunruamkaew K. How to do AHP analysis in Excel. University of Tsukuba, Graduate School of Life and Environmental Sciences, Division of Spatial Information Science. 2012.


Appendix A: Consistency Analysis

The consistency analysis was carried out to check the consistency of a single interviewee's responses on each dimension (20, 49-51, 55). See Figures A.

Figure A: An Example of Consistency Calculation (20, 51)

In cases of exceptional inconsistency, we went back to the assessor for clarification and corrected the responses accordingly. There were inconsistencies in 4-6 questions for each interviewee (out of 138 questions).
Appendix B: Calculation of Geometric Prioritization RGMM Including Consensus Indicator

A. Calculation of geometric prioritization by the row-wise geometric mean method (RGGM) for each participant. This was done for all the diagnoses in all dimensions depending on the interviewee's answer (55). See Figure B.1-B.2.

1. Calculating ln function for the values in the Comparison Matrix (Figure B.1)

Figure B.1: Ln Matrix Calculation

\[
\text{Ln(A) for calculation of RGMM} = \begin{pmatrix}
1 & -1.39 & -0.69 & 1.10 & -0.69 & -1.39 & 1.10 & -0.69 & -0.69 \\
2 & -1.39 & -1.39 & -0.69 & -1.39 & -1.79 & -1.10 & -1.39 & -1.79 & -1.39 \\
3 & 0.69 & 1.39 & -1.10 & 1.10 & 1.10 & -1.39 & 1.10 & -1.39 & -0.69 & -0.69 \\
4 & -1.10 & 0.69 & -1.10 & -1.10 & -1.10 & -1.79 & -1.10 & -1.39 & -1.79 & -1.39 \\
5 & 0.69 & 1.39 & -1.10 & 1.10 & -1.10 & 1.10 & -1.39 & -0.69 & -1.39 & -0.69 \\
6 & 1.39 & 1.79 & 1.39 & 1.79 & 1.10 & -1.39 & 1.10 & 0.69 & 1.10 \\
7 & -1.10 & 1.10 & -1.10 & 1.10 & -1.10 & -1.39 & -1.10 & -1.39 & -1.10 \\
8 & 0.69 & 1.39 & -1.39 & 0.69 & 0.69 & -1.10 & 1.10 & -1.10 & -0.69 \\
9 & 0.69 & 1.79 & 0.69 & 1.79 & 1.39 & -0.69 & 1.39 & 1.10 & -1.10 \\
10 & -1.39 & 0.69 & 1.39 & 0.69 & -1.10 & 1.10 & 0.69 & -1.10
\end{pmatrix}
\]

2. Using EXP function to calculate the value of the mathematical constant e, raised to the power of the average for each row in the matrix (Figure B.2).

3. Calculating the RGMM by dividing the EXP result of each row in the summary of all rows (Figure B.2).
B. Then after achieving consistency as illustrated in Appendix A, a weighted calculation was made of the pairwise comparisons of all respondents across the ICD-9 diagnoses for each dimension as described in (56). See Figure B.3-B.5.

The consolidated calculation was the weighted geometric mean of all the participants.

1. The consolidated decision matrix combines all participants’ inputs to obtain the aggregated group result. We used the weighted geometric mean of the decision matrix's elements $a_{ij}$ using the individual decision maker’s weight as given in the input sheets:

$$
\bar{x} = \left( \prod_{i=1}^{n} x_i^{w_i} \right)^{1/\sum_{i=1}^{n} w_i} = \exp \left( \frac{\sum_{i=1}^{n} w_i \ln x_i}{\sum_{i=1}^{n} w_i} \right)
$$

2. Matrix multiplication iterations. We stop repeating when the difference between the sums in the two consecutive calculations was smaller than a prescribed value.

3. Scaling-finding the corresponding weight by dividing each value by the max in the same column.
4. Normalization: dividing each element in the matrix by its column total to generate a normalized pair size matrix.

5. Calculating the weighted average by multiplying the sum of each column in the basic matrix by the normalized weight.

Figure B.3: The weighted calculation of all participants
C. Checking consistency across all respondents (49, 55). We applied the stringent approach to inconsistencies in (39) who recommended that a deviation of over 10% in the inconsistency coefficient constituted an inconsistency. The results in our case were lower than this value. The AHP consensus was calculated based on the RGMM results of all inputs using Shannon alpha and beta entropy (51). The consensus indicator ranges from 0% (no consensus between decision makers) to 100% (full consensus between decision makers). See Figure B.6.
\[ S = \frac{M - \exp(H_{\alpha, \beta, \gamma})}{\exp(H_{\alpha, \beta, \gamma}) / \exp(H_{\alpha, \beta, \gamma}) / (1 - \exp(H_{\alpha, \beta, \gamma}) / \exp(H_{\alpha, \beta, \gamma}))} \]

With \( M = 1 / \exp(H_s) \)

\( H_{\alpha, \beta, \gamma} \) is the \( \alpha, \beta, \gamma \) Shannon entropy for the priorities of all \( Z \) decision makers/participants (55)

Shannon alpha (see stage 4 in figure B.6 below):

\[ H_\alpha = \frac{1}{Z} \sum_{j=1}^{Z} \sum_{i=1}^{N} p_{ij} \ln p_{ij} \]

With \( \bar{p}_j = \frac{1}{Z} \sum_{i=1}^{N} p_{ij} \)

Shannon gamma (see stage 5 in figure B.6 below):

\[ H_\gamma = \sum_{j=1}^{Z} - \bar{p}_j \ln \bar{p}_j \]

Shannon beta (see stage 6 in the calculation in figure B.6 below):

\[ H_\beta = H_\gamma - H_\alpha \]

With \( H_{\alpha, \min} = - \frac{M}{N+M-1} \ln \left( \frac{M}{N+M-1} \right) - \frac{1}{N+M-1} \ln \frac{1}{N+M-1} \)

And \( H_{\gamma, \max} = (N - Z) \left( \frac{1}{N+M} \ln \frac{1}{N+M} + \frac{Z-1}{M+Z-2} \right) \ln \frac{1}{M+Z-2} \ln \frac{1}{Z-1} \)

\( N \) number of classes, \( Z \) number of samples, \( M \) max. Score of AHP scale (lin=9)

Figure B.6: Consensus Indicator based on RGGM results.