A New Paradigm to Analyze Data Completeness of Patient Data

Ayan Nasir[1]; Varadraj Gurupur[1]; Xinliang Liu[1]
[1]Department of Health Management and Informatics, University of Central Florida

Keywords
Concept maps, CSV parsing, data completeness, electronic medical/health record

Summary
Background: There is a need to develop a tool that will measure data completeness of patient records using sophisticated statistical metrics. Patient data integrity is important in providing timely and appropriate care. Completeness is an important step, with an emphasis on understanding the complex relationships between data fields and their relative importance in delivering care. This tool will not only help understand where data problems are but also help uncover the underlying issues behind them.
Objectives: Develop a tool that can be used alongside a variety of health care database software packages to determine the completeness of individual patient records as well as aggregate patient records across health care centers and subpopulations.
Methods: The methodology of this project is encapsulated within the Data Completeness Analysis Package (DCAP) tool, with the major components including concept mapping, CSV parsing, and statistical analysis.
Results: The results from testing DCAP with Healthcare Cost and Utilization Project (HCUP) State Inpatient Database (SID) data show that this tool is successful in identifying relative data completeness at the patient, subpopulation, and database levels. These results also solidify a need for further analysis and call for hypothesis driven research to find underlying causes for data incompleteness.
Conclusion: DCAP examines patient records and generates statistics that can be used to determine the completeness of individual patient data as well as the general thoroughness of record keeping in a medical database. DCAP uses a component that is customized to the settings of the software package used for storing patient data as well as a Comma Separated Values (CSV) file parser to determine the appropriate measurements. DCAP itself is assessed through a proof of concept exercise using hypothetical data as well as available HCUP SID patient data.

Correspondence to:
Varadraj Gurupur
Department of Health Management and Informatics
University of Central Florida
Email: varadraj.gurupur@ucf.edu,
varadrajprabhu@gmail.com

Appl Clin Inform 2016; 7: 745–764
http://dx.doi.org/10.4338/ACI-2016-04-RA-0063
received: April 26, 2016
accepted: July, 4 2016
published: August 3, 2016
Citation: Nasir A, Gurupur V, Liu X. A new paradigm to analyze data completeness of patient data. Appl Clin Inform 2016; 7: 745–764
http://dx.doi.org/10.4338/ACI-2016-04-RA-0063

This document was downloaded for personal use only. Unauthorized distribution is strictly prohibited.

© Schattauer 2016

1. Background and Significance

As health care moves into the 21st century, practitioners are moving into an era where traditionally qualitative methods of care are being complemented with quantitative research and analysis to better understand and solve various problems [1-5]. To that end, one important problem health care practices currently face is the strength of their patient records and databases. From basic information such as phone numbers or email addresses to more involved information such as family history and procedure records, there is no centralized mechanism through which data entered by health care staff can be checked for completeness and validity [4]. Furthermore, there are no centralized mechanisms through which various health care software database packages and health center record-keeping measures can be compared to determine advantages and disadvantages of different systems [4]. This decentralization also has tangible effects on patient care (such as having to access outside clinical information to fill in gaps in information/verify patient information [5] or the actual types of information requested [6]), often times leading to issues in timeliness of information or the accuracy of information in leading to improper care or treatment plans. Without having a mechanism that can identify strengths and shortcomings in record-keeping procedures, it becomes more difficult to ensure the highest standard of care across health care centers. Most importantly, these problems are being uncovered at a time where data accuracy in health care is becoming a cornerstone for the burgeoning field of health informatics as well as a fundamental part of new models of health care reimbursement and practice [7]. With the importance of complete patient data in mind, this paper sets out to provide a proof-of-concept study in developing a tool that can intelligently determine patient data completeness. Current systems in place either do not address patient data completeness or assess it in a very simplistic manner while this paper strives to ensure that completeness is represented in the context of the relative importance of data fields. Alongside developing this tool, the paper sets to develop a framework upon along subsequent clinically relevant research can be conducted, such as analyzing underlying causes for incomplete data and specific subpopulations that may be at risk of incomplete data.

From a foundational level, significant work has been done on understanding electronic medical records, their advantages and disadvantages, and the effects they have in the health care environment [8, 9]. For example, previous research has focused on topics such as providing clinician's perspective on electronic health care records [10], how health care records developed and became connected across health care centers [11], and how health care records became data sources for hypothesis driven research projects [12]. One research paper highlights the general problem of accuracy in computer-based patient records (CPR), concluding that their „... knowledge of data accuracy in CPRs is not commensurate with its importance and further studies are needed" [13]. They provide an understanding of the mechanism involved in the actual entry of the data (▶Figure 1 provides an updated schematic detailing this mechanism and the various personnel involved) as well as a basic mathematical analysis of completeness and correctness. Nonetheless, their work focused more on identifying problems in CPRs and did not lead to the development of an automated tool. Another research paper [14] however did implement a quality assessment tool that focused on accuracy of data in a specific primary care university clinic system and noted a reduction in errors when implementing the quality assessment tool. The primary focus of this work is to help personnel with errors in database entry and produce a data quality assessment tool limited to the health center being evaluated. The creation of a framework [15] in evaluating patient data strength (through completeness, correctness, concordance, plausibility, and currency) was an important development in laying the foundation for understanding how to evaluate data. However, this work was limited to literature survey and did not create a program through which to assess current patient data on these axes.

Other research projects further investigate this problem as it pertains to primary care record keeping accuracy and completeness [16], noting that completeness and accuracy of data in primary care is highly dependent on how receptive and understanding practitioners are about the importance of patient data. They also explain how there is no central reference standard for data quality for primary care patient records, which limits their effectiveness. While providers generally understand the importance of data, obstacles such as cost and time limit positive change. Furthermore, there remains a lack of important standards in data exchange between healthcare standards that causes issue in patient record accuracy and care timeliness. This paper is important in not only highlighting the

© Schattauer 2016

This document was downloaded for personal use only. Unauthorized distribution is strictly prohibited.

issue of data inaccuracy and incompleteness but also noting some underlying issues and consequences (specifically the lack of potential benefits).

Along with this paper, several studies in different specialties and different geographical areas have examined the problem of incomplete and inaccurate data within specific databases [17, 18]. There has also been significant work in survey and literature review of databases to understand underlying issues on why data may be incomplete or inaccurate [19-21]. From a mathematical and statistical perspective, understanding how to properly analyze numerical data in a medical context is an area that has been heavily explored [22, 23]. There has also been significant analysis from a database and computer systems perspective, specifically on how data is entered, issues with interface, and sources of errors that stem from the connection between personnel and computer systems [24, 25]. Furthermore, research focused on mining clinical data [26] and understanding how to develop robust health care technology information systems [27] are vital areas of interest in developing strong patient databases that can be used for better delivery of care.

While these papers lay the groundwork for understanding the problem, they do little in terms of providing solutions that could be widely implemented to ensure data completeness. For a clearer understanding of this objective and understanding of current solutions, ▶Table 1 is used to compare the tool generated in this project with Patient-Centered Medical Home (PCMH) quantitative record keeping benchmark methodology [28-30] and private industry solutions [31-33].

Lanzola et al. [34] provide work in an important related area. While this paper focuses on developing a system to analyze databases in their current states, this study proposes a 're-engineering process' that improves data completeness and accuracy (which in turn provides stronger data from which quality of care and information mining can occur). This study highlights areas of concerns, which are covered in areas of further exploration, and provides a foundation for issues that this research project addresses from the perspective of existing systems.

2. Objectives

Based on the aforementioned description of the current state of research in this area, a project was developed with the following objectives:

- Develop a tool that can determine the completeness of individual patient records as well as aggregate patient records across health care centers and subpopulations
- Equip the tool with robust statistical analysis to ensure that the importance of the various types of data is accounted for
- Design the tool to allow health care staff to evaluate the strength of record keeping and identify areas for improvement.

3. Methods

Focused on the objectives outlined in the Introduction, a tool is developed that analyzes health care data and complements concept mapping as a way to thoroughly represent data stored. This tool is termed the Data Completeness Analysis Package (DCAP).

3.1 Concept Mapping and Data Handling

One major premise to the approach invoked in this project is to use concept maps as a representation of patient record data. Concept maps are visual representations of data stored in a system, traditionally used for expressing knowledge in both a declarative and procedural sense [35]. Most concept mapping tools have been previously used to express knowledge in a teaching environment to evaluate student learning of various topics [36, 37]. An example is provided in ▶Figure 2. For the purposes of this project however, concept mapping is used as a schema through which patient data could be represented and uniformly examined independent of the platform the data was originally stored on or other health care protocols that may make cross-examinations of data sets more difficult. These concept maps are developed manually and then converted to CSV files, which can then

This document was downloaded for personal use only. Unauthorized distribution is strictly prohibited.

be analyzed in DCAP through a parser that allows the user to determine the strength of both individual patient records as well as the strength of record keeping throughout the database. For this project, IHMC (Institute for Human and Machine Cognition) CMapTools is used to create concept maps.

This patient concept map is generated using data stored in the health care database and mapped using a user-developed ‚master map'. This master map is a schema that represents all the information (both in terms of the data fields themselves and the proper interconnectedness amongst those data fields) the way the user wishes them to be stored in order to be considered ‚complete'. While within a health care center this master map may be constrained by the database software used or methods of patient data recording used by staff, when comparing data sets from various health care centers using DCAP the master map can be broadened as necessary. An example of both this master map and a sample patient map (▶Figure 3) are shown below to highlight how the concept of completeness (and subsequently strength of individual/aggregate patient data) can be visualized. While these examples are not comprehensive, they show how uniform concept map representations serve as a platform through which basic shortcomings in patient data can be understood. It is also important to note that while cross subgroup relationships and other schema complexity may be useful in patient record applications (for example, address and phone number are concepts that can be linked to both personal information and billing information), for the purposes of this project having the information as defined by the master map is sufficient to determine completeness. Since this research project is a proof-of-concept exercise, the master map samples are for illustrative purposes created by the author to explain their usage. The concept maps are also generated manually upon analyzing the database and variables used.

Once the native patient data in the health care database is converted into concept map representations using a template master map, the concept maps are then converted into CSV documents. ▶Figure 4 shows a sample CSV document which includes the concept labels (i.e. patient data fields). A CSV analyzer then parses the documents to see which fields are filled/unfilled and then sends its results for statistical analysis. DCAP is flexible in assigning various conditions for completeness and can be altered for individual data fields and different subpopulations or databases. After the relevant statistics have been generated (e.g. individual patient records, entire database completeness, etc.), a text file is generated to display the results for the user. ▶Figure 5 provides a flowchart for a better understanding of how DCAP works.

3.2 Data Analysis

Once the CSV analyzer has processed the patient data, DCAP is able to implement statistical analysis in order to determine the completeness and strength of the patient data. While completeness in each individual data field is easily identifiable by the presence (or lack thereof) of data, analyzing an individual patient's record holistically (or an entire database or subgroup of patients) requires determining which data fields are more important than others. With this in mind, DCAP generates a Record Strength Score (RSS) that is based upon the users input of Importance Weights (IW).

To illustrate this concept, two examples are used: one where each data field holds an equivalent importance (and thus equivalent IWs), and another where the IWs are different. For the scoring itself, each IW ranges from 0 to 100 points, with 100 denoting maximum importance. For a record with all data fields of equivalent importance, 100 points are used for each data field. For a record with data fields of various levels of importance, the IW is adjusted along the scale accordingly. It is also important to note that while the concept map shows various grouping concepts that organizes the data fields (such as ‚personal information', ‚insurance', or ‚medical background'), these groupings have no scoring. Thus the only scoring concepts are those with actual corresponding data fields. Furthermore, note that setting an IW to 0 points effectively makes that field optional. The scoring equation is as follows:

$$RSS\ (\%) = \frac{(IW_1)(X_1) + (IW_2)(X_2) + (IW_3)(X_3) + \ldots + (IW_n)(X_n)}{\sum_{i=1}^{n} IW_i} \times 100 \quad (1)$$

© Schattauer 2016

This document was downloaded for personal use only. Unauthorized distribution is strictly prohibited.

Where

RSS = Record Strength Score (Total Strength of Completeness for Individual Patient Record constrained from 0 to 100%)
IW_i = Importance Weight of 'i'th data field
X_i = Binary Completeness Variable for 'i'th data field (1 represents complete data field, 0 represents incomplete)

▶Figure 6 shows the sample patient master map for balanced scoring. ▶Figure 7 follows with the sample patient map indicating the Binary Completeness Variable for each field. Based on those maps and the scoring equation, the RSS for the sample patient is 64%.

▶Table 2 shows IWs for an unbalanced RSS, and ▶Figure 8 shows the new sample patient master map with modified IWs (the Binary Completeness Variables for the sample patient shown in ▶Figure 8 remain unchanged). Please note that the weights used are for purposes of example and have no bearing on actual or perceived relative importance of data fields. Based on those maps and the scoring equation, the RSS for the sample patient is 58%, showing that these new IWs make this patient record ‚less complete'.

While there is no certain answer for an optimal scoring algorithm that can take into account various preferences for the importance of individual patient record data fields, implementing the RSS scoring equation in DCAP allows for the score to be tailored to the needs of the user. Further development and use of DCAP can also lead to discoveries in better scoring algorithms, and the possibilities with those improvements are elaborated upon in the ‚areas of further exploration' section of this paper.

Another important metric to establish is the Patient Database Score (PDS), which defines the overall strength of all records in the database and is simply the average of the various RSS scores (Equation 2). This metric allows for comparisons between databases to determine better record keeping software packages and strategies. Furthermore, using database segmentation techniques, subpopulations of patient records can be compared using the Patient Subgroup Score (PSS), which averages the RSS scores of the patients of interest (ex. by age, race, gender, insurance status, etc.) and can allow for an in depth analysis of record strength based on patient information (Equation 3). This topic is further elaborated upon in the ‚areas of further exploration' portion of this paper.

$$PDS\ (\%) = \frac{RSS_1 + RSS_2 + \cdots + RSS_n}{n} \quad (2)$$

Where

PDS = Patient Database Score (Average of all patients RSS within the database)
RSS_i = Record Strength Score (Total Strength of Completeness for Individual Patient Record constrained from 0 to 100%) for ith patient
n = Total number of patient records

$$PSS\ (\%) = \frac{RSS_1 + RSS_2 + \cdots + RSS_n}{n} \quad (3)$$

Where

PSS = Patient Subgroup Score (Average of all patients RSS within the database ***that meet subpopulation condition)***
RSS_i = Record Strength Score (Total Strength of Completeness for Individual Patient Record constrained from 0 to 100%) for ith patient ***that meets subpopulation condition***
n = Total number of patient records ***that meet subpopulation condition***

Individual data fields across the patient database can also be compared using the Data Field Completeness Score (DFCS), which averages the binary completeness variable of a specific data field across all (Equation 4) (or a subgroup, Equation 5) of patients to determine how well one particular class of data is being recorded. The DFCS allows for comparisons across databases and subgroups to

© Schattauer 2016

This document was downloaded for personal use only. Unauthorized distribution is strictly prohibited.

ensure proper patient record keeping for each data field. The various scores and terms of importance in scoring are summarized in ▶Table 3.

$$DFCS_k\ (\%) = \frac{\sum_1^n X_{ik}}{n} \times 100 \quad (4)$$

Where

DFCS = Data Field Completeness Score (Percentage completeness of a certain data field among all patients within the database)
X_{ik} = Binary Completeness Variable for 'i'th patient (1 represents complete data field, 0 represents incomplete) and kth data field, where k remains constant and i iterates for individual patients
n = Total number of patient records

$$DFCS_k\ (\%) = \frac{\sum_1^n X_{ik}}{n} \times 100 \quad (5)$$

Where

DFCS = Data Field Completeness Score (Percentage completeness of a certain data field among all patients within the database ***that meet the subpopulation condition****)*
X_{ik} = Binary Completeness Variable for 'i'th patient (1 represents complete data field, 0 represents incomplete) and kth data field, where k remains constant and i iterates for individual patients ***that meet the subpopulation condition***
n = Total number of patient records ***that meet subpopulation condition***

4. Results

4.1 Testing DCAP through Hypothetical Data

In order to build and test DCAP, first a random population set was generated in Excel using the same fields as shown in ▶Figure 3 and ▶Figure 4. The Excel random function was used to introduce random missing fields and the file was converted into a CSV format to be used by DCAP. In terms of specific mechanics of DCAP itself, it reads in three CSV files:

1. Patient File – These are each of the patients records in one CSV, with columns representing fields and rows representing individual patients
2. Template File – These are all of the field labels (one row with multiple columns matching in order to the fields stored in the patient file)
3. Importance Weight File – These are the importance weight corresponding to the template file

DCAP then takes these files, converts them to arrays, checks for completeness (against a defined string variable that indicates incompleteness which can range from a first character indicator to a blank value), and then outputs the general result of the PDS for the entire database and RSS for each patient, with each of these metrics also calculated for balanced IWs (so that one can compare a simple average to the average weighted by user IWs in the Importance Weight file). DCAP also automatically generates the DFCS for each field. ▶Figure 9 shows a sample result. Upon request of the user, DCAP further generates PSS and subpopulation DFCS based on either numerical bound or text bound data field subpopulation conditions. ▶Figure 10 provides a sample result for the hypothetical data where PSS and subgroup DFCS were generated (the subpopulation were patients between the age of 0 and 50 and the DFCS field was race).

Testing with hypothetical data served to ensure that DCAP provided accurate results and that the framework to analyze patient data completeness was robust. After verifying its accuracy and usage (via manual analysis in Excel), DCAP could be employed on real data sets. Hypothetical data also allowed for DCAP speed and memory usage to be tested. At up to ~1000 patients, the program could run within a few seconds without significant memory usage. When increasing patient size to 100,000, the program required additional virtual memory (setting maximum memory size to 3072

This document was downloaded for personal use only. Unauthorized distribution is strictly prohibited.

Mb) and took closer to a minute to run. Running the program at a million records and above caused DCAP to crash, showing an important limitation to the current record capacity of DCAP.

4.2 Experimenting with HCUP Data

In order to test DCAP on real patient data, de-identified data from the Healthcare Cost and Utilization Project (HCUP) [38] was used, specifically the State Inpatient Database (SID) [39] with data from Florida. In setting the importance weights, it was noted that many fields were used to handle excess information (for example multiple diagnosis codes – there are 31 fields for up to 31 diagnoses, but having a blank in those fields was not actually resultant of missing information and thus IWs were adjusted to 0). Otherwise, importance weights were set first to balanced and then based on general view of importance. Due to the large number of data variables contained in the HCUP data, the IWs have been omitted but can be presented by the corresponding author to interested readers.

▶Figure 11 and ▶Figure 12 show the results of the HCUP data in 2012 for the balanced IWs and unbalanced IWs respectively. Please note that the PDS generated with 'balanced IWs' in the result print out incorporates fields as mentioned above such as multiple diagnosis codes and thus are not a good measure of record completeness. This simple analysis of HCUP SID data in Florida shows both that data is not complete to a level necessary for proper recording of healthcare service and that when the importance of various fields are taken into consideration that measure becomes even weaker. This data and the functionality provided by DCAP allows for multiple areas of further exploration that are discussed later in this paper.

After using DCAP on the HCUP SID data for each year, ▶Figure 13 shows changes over time in PDS scores for the samples. While the unbalanced IWs that were inputted by the authors are for illustrative purposes only, it is vital to note the gap between when IWs are balanced (i.e. each data field has equal importance) and when they are not, pointing towards further research in discriminating between the importance of different data fields. It is also necessary to contextualize both the gap in PDS between balanced and unbalanced IWs and gaps in PDS over time by evaluating standard deviations. As ▶Figure 14 shows, the standard deviation for unbalanced user IWs is consistently higher. It also shows that during periods of lower PDS, standard deviation of the sample rises. Combining these findings, ▶Figure 15 shows PDS over time in context of one standard deviation of error, minimizing discrepancies over time but still showing persistence in gap between balanced and user IWs. Also important to note (as shown in Appendix A.2) is that virtually all deviations indicated in the line graph between PDS scores over times are of great statistical significance as calculated by the two-sided t test. This test is suitable based on statistical literature [40, 41] in that the baseline assumptions needed (sufficiently large sample size and normality) are present, with the null hypothesis denoting no statistically significant change in PDS over time and the alternative hypothesis showing a statistically significant change in PDS over time (in either direction). This further highlights the importance of discrepancies in PDS, PSS, and RSS scores and provides an avenue for hypothesis-driven analysis as discussed in the subsequent section (for example, is there a trend persisting in data completeness over time? Are certain subpopulations at greater risk for incomplete patient data?). These questions can now be explored through the framework of analysis and data handling developed through DCAP and addressed in areas for further exploration.

4.3 Performing Validation Tests

In order to verify the results that DCAP provided with regards to the HCUP data, 50 random patients were selected from the original data spanning all the years used. These records were then manually analyzed to compute their RSS scores (for both balanced IWs and unbalanced IWs) and then compared to the results provided by DCAP. These records were also used to calculate a standard percentage completeness (showing what percent of data fields were filled regardless of the type of data). The results of this verification points to two important conclusions. The first is that DCAP is accurate in calculating RSS scores, with either no error between manually calculated scores and DCAP scores, or marginal errors resulting from truncation during calculation. Secondly, the comparison of scores to a simple percentage basis shows a wide rift between the percentage calculation and both the balanced and unbalanced IW RSS scores. This shows the importance of accounting for

This document was downloaded for personal use only. Unauthorized distribution is strictly prohibited.

the relative importance in different data fields to obtain a true understanding of how complete patient data is.

5. Discussion

5.1 Limitations

During the course of this project, various important limitations were encountered that we hope can be addressed in subsequent work on the topic of health care record strength.

First, based on our experience with this project, excising data from actual health care service centers and putting it into a uniform file across various database software packages is a difficult task fraught with obstacles. The willingness of health care service centers is important, with patient confidentiality and Health Insurance Portability and Accountability Act (HIPAA) regulations limiting the ability of health care centers to provide data. Furthermore, the limitations of staff at these centers also prove difficult in obtaining data (since very few private practices have on site data technicians able to export fully de-identified and HIPAA compliant data). This leads into another problem of clinical data software packages and incentives of data export. For many practices, exporting data for any reason is difficult based on the parameters and complexity of these software packages. This often requires either the software vendor itself or a third party IT consultant to be involved and makes it difficult to put the data into usable form. It should also be noted that even when data is obtained, such as the HCUP data, there is still some degree of manual data handling to make it compatible for use with DCAP (ex. correct file format, understanding what an incomplete variable means per data element, etc.).

Another limitation is the inability of DCAP currently to calculate the veracity, and not just the completeness, of health care data. While knowing the completeness of data is important, knowing that the data is legitimate and reliable is of great use to any health care provider. This also leads into the issue of free text fields. In many health care data variables, practitioners implement free text notation to account for the variety of conditions and treatments noted, which may cause issues in determining veracity. For the purposes of this project, the data in HCUP SID was stored in simple variable form and completeness was identifiable. Furthermore, free text data fields present less issue with completeness in that any indicated notation would imply that the field was filled (but makes no presumption on its accuracy).

From a concept mapping perspective, currently the concept maps are created manually in a separate program and are used strictly as a schema to represent data. Due to the manual nature of conversion from native database to concept map, there are still limitations with regards to setting up DCAP for different data sets as well as ease-of-access for health care staff that may not necessarily understand their data storage and protocols. For the purposes of this project, concept mapping is used to represent the data storage schema. In eventual application, an automated procedure would allow for the schema to be determined without manual analysis and patient files would be more robust in their representation of data with regards to hierarchies and linkages. Outside of concept mapping, DCAP has limitations in terms of memory needed to run for larger data sets and its inability to handle extremely large patient sample sizes (1 million or more at a time).

As in any statistical analysis, there are limitations to exactly what one can gleam from the statistics generated. For example, the determination of relative importance weights is subjective upon the unique objectives of each user of DCAP and thus various scores have various levels of applicability to different users. Furthermore, there are also limitations to DCAP that one can analyze and expand in later revisions. For example, the ability to combine subpopulations to generate PSS and subgroup DFCS scores (▶Table 3) or a more robust subpopulation parameter system would allow DCAP to be more intelligent in establishing important links between subpopulations and record completeness. Another limitation on the software end is the lack of automation in concept mapping of patient records, which makes data visualization a manual process (and a cumbersome one for complex and thorough patient records).

This document was downloaded for personal use only. Unauthorized distribution is strictly prohibited.

5.2 Areas of Further Exploration

These limitations discussed above lead in to areas of further exploration that subsequent research might delve to learn more about. For example:

- More work can be done in the area of subpopulation completeness, and public databases such as the HCUP data provide ample data to answer questions of health care record keeping as it pertains to vulnerable populations.
- Develop and test DCAP further for use alongside a variety of health care database software packages in current active health care settings
- Further research and analysis should also be conducted in automating concept map building [42] within DCAP for users and health care personnel to better visualize data and understand the relative importance of different data elements. Furthermore, future studies could also focus on detailed clinical data that cannot be codified in singular data elements to understand how this data can be assessed for completeness and validity.
- Analyze use of data standardization at both patient record level and individual data element level as ways to assess and ensure completeness/validity.
- Patient record strength, completeness, and veracity can ultimately be linked to other issues such as various diagnoses, rates of injury, mortality, etc.
- Focusing on information integrity directly related to diagnoses, prescriptions, and incongruences between the two
- Further explore statistical and mathematical issues that develop, such as determining what constitutes a significant difference in completeness between patients, subpopulations, and databases
- The issue of automated veracity could be investigated through common data heuristics or through communication with a central database as a cross-reference. Special attention needs to be paid to free-text data variables (for example, the MetaMap tool project [43] that is being used to index free biomedical text).
- IWs can be made less subjective and more objective based on a panel of health professionals to ensure that the completeness of a database is an objective measure based on widely agreed upon standards.
- Focusing on solutions involving using new software packages and database standards that integrate lessons learned in maintaining complete and accurate data [34]. This includes topics such as: data entry control, use of contextual information, incentive alignment, user requirement analysis, measurement of improvement, data entry checks, enhanced data analysis, and creating strong networks among disparate databases to cross-check information. Developing a user-interface that can facilitate a wide range of data entry while being broadly accepted by health care providers is also critical in furthering patient data integrity, analysis, and management.
- DCAP could be improved in terms of handling larger data sets with greater efficiency, both in terms of speed and memory usage. For example, the hypothetical data was used in this paper allowed DCAP to be stress-tested and database size limitations to be found by increasing patient sample sizes until the program would crash or run out of memory. Future work could focus on creating more efficient algorithms or focusing on the computing requirements necessary to ensure the verification of complete and accurate data.

6. Conclusion

The core contributions of this project are as follows:

- The development of a system of mathematical equations to express the completeness of both individual patient records and database/subgroup patient records
- The development of a tool that assesses the data completeness of patient records based on the aforementioned framework
- The development of a framework to understand problems regarding patient data completeness, accuracy, and interrelatedness that provides a foundation for future work in the area

This document was downloaded for personal use only. Unauthorized distribution is strictly prohibited.

Based on the original objectives, the tool created was successful in applying the statistical metrics evaluated in this paper and able to evaluate the strength of a health care database (specifically, the HCUP SID Florida database). It also demonstrated ease of use and wide accessibility (interested readers can consult the appendix for a video demonstration of the program). Limitations in access to current clinical data stored in proprietary software packages led to a lack in testing against various healthcare databases but the tool provides promise in its adaptability to various data sets and data elements. Overall, this research project uses real data to show the current problem of incomplete patient data and the importance of creating systems to ensure a high standard of data completeness. In its development and limitations it also builds a foundation to explore further areas of interest and advocates work committed to making sure that important patient data is complete and accurate.

7. Clinical Relevance Statement

The creation of DCAP is important in clinical practice in that it helps identifies issues in patient data completeness. Current systems in place either do not address patient data completeness or assess it in a very simplistic manner while DCAP ensures that completeness is represented in the context of the relative importance of data fields. The development and testing of DCAP has provided a platform upon which subsequent clinically relevant research can be conducted, such as analyzing underlying causes for incomplete data and specific subpopulations that may be at risk of incomplete data.

Involvement of Human Subjects

The investigators have conducted the experiment after getting appropriate approval from the Institutional Review Board (IRB) of the institution (IRB Exempt). Therefore all the required ethics and standards have been followed.

Conflict of Interest Statement

The authors have no conflicts of interests to report.

Acknowledgements

We would like to thank Dr. Thomas Wan for his valuable guidance on this project. We would like to inform the readers that this article is an extended version of the paper published in the Proceedings of SDPS 2015 Annual Conference published by the authors and permission has been granted by the society to publish it in any journal.

This document was downloaded for personal use only. Unauthorized distribution is strictly prohibited.

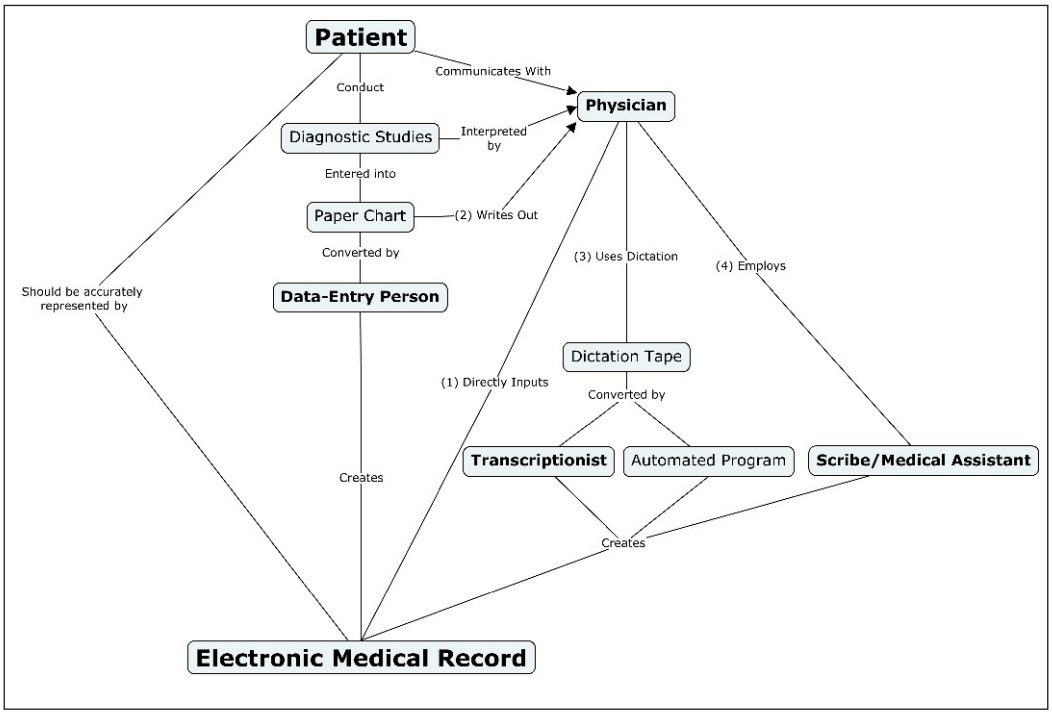

Fig. 1 Patient record entry flowchart adapted and updated from Hogan and Wagner

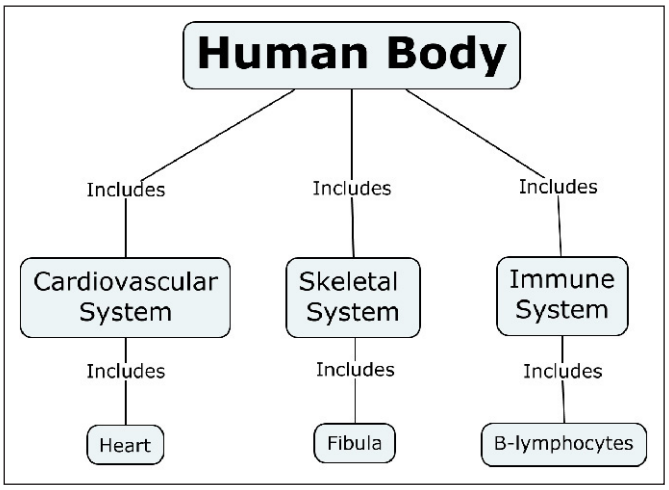

Fig. 2
Concept Mapping Example in Education; Student's understanding of the human body

This document was downloaded for personal use only. Unauthorized distribution is strictly prohibited.

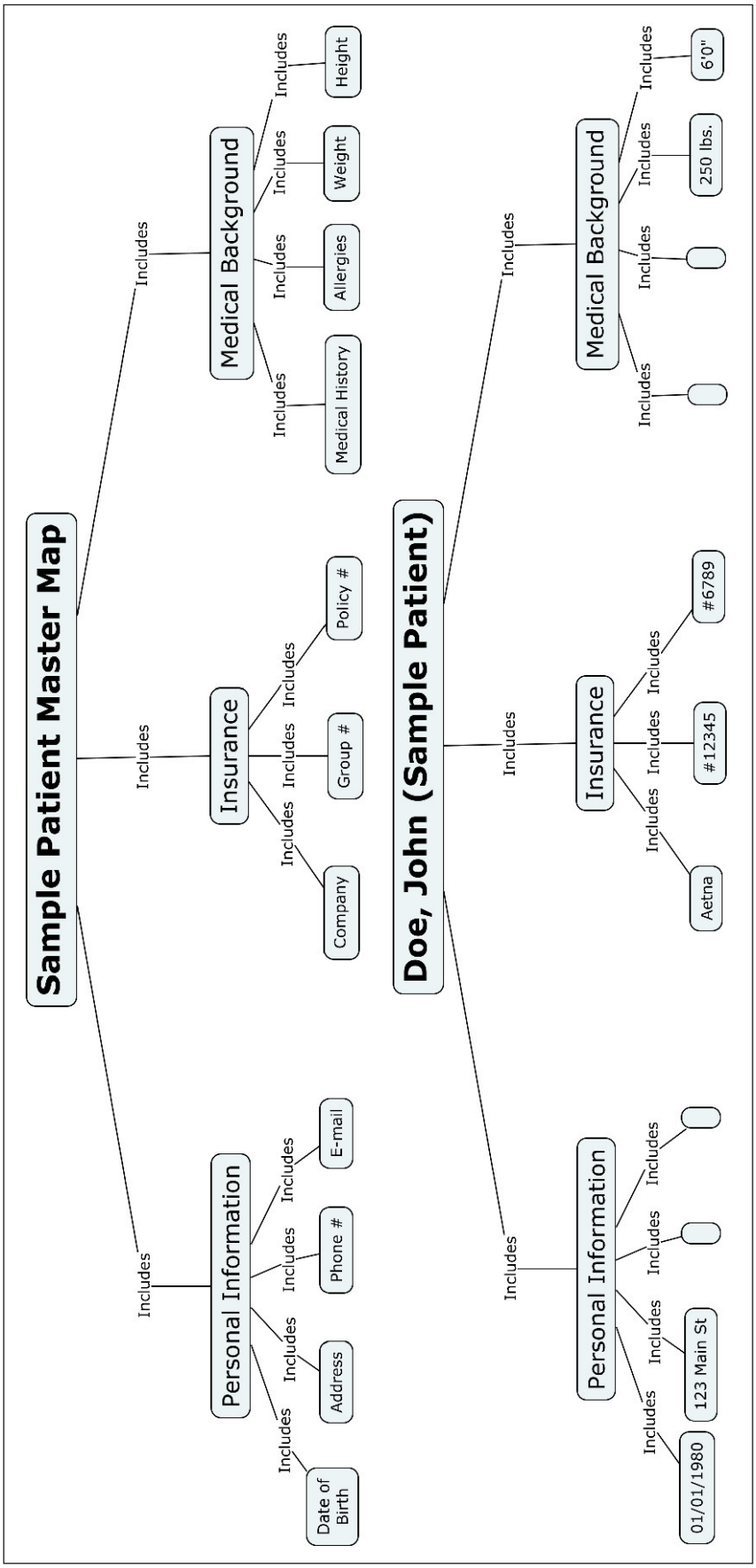

Fig. 3 Sample Patient Concept Master Map

This document was downloaded for personal use only. Unauthorized distribution is strictly prohibited.

Template

```
Last Name,First Name,DoB,Address,Phone #,Email,Ins Co,Ins Group #,Ins Policy #,Med
History,Allergies,Weight,Height,Age,Race
```

PatientFiles

```
P1,P1,2/22/54,Complete,8946217934,Complete,N/A,N/A,8495985914,Complete,N/A,292,77,57,Blue
P2,P2,7/17/64,Complete,9629286899,Complete,N/A,3354734608,9601255657,Complete,Complete,
315,56,47,N/A
P3,P3,6/16/85,Complete,9832107225,N/A,N/A,4085697516,N/A,Complete,Complete,214,84,26,Red
P4,P4,N/A,N/A,2486894508,Complete,Complete,N/A,9838931688,Complete,Complete,149,N/A,N/
A,Blue
P5,P5,N/A,Complete,8229701469,Complete,Complete,1970108556,1450221526,Complete,Complete,
135,N/A,N/A,N/A
P6,P6,6/18/86,Complete,9993459610,N/A,N/A,8932757385,6397254206,Complete,Complete,
107,62,25,N/A
P7,P7,6/14/09,Complete,1857558413,Complete,N/A,5774440050,N/A,N/A,N/A,135,N/A,2,N/A
P8,P8,N/A,Complete,1098295438,N/A,Complete,8561178035,7981797488,Complete,Complete,236,N/
A,N/A,Green
P9,P9,3/27/36,Complete,2140491160,N/A,Complete,7901815542,2054290177,Complete,Complete,
294,70,75,N/A
P10,P10,N/A,Complete,7198293887,Complete,Complete,7110963024,8140505368,Complete,Complete,
274,83,N/A,Red
P11,P11,11/26/04,Complete,N/A,Complete,N/A,7505898530,7371637729,Complete,Complete,
327,59,7,N/A
P12,P12,1/2/02,Complete,6936339182,Complete,Complete,1723365281,7362707730,Complete,N/A,
270,N/A,9,Orange
P13,P13,11/1/79,N/A,N/A,Complete,Complete,2082387399,N/A,Complete,Complete,
114,70,32,Violet
P14,P14,9/29/96,Complete,7558942144,Complete,Complete,8966736543,1589144199,N/A,Complete,
322,N/A,15,Red
P15,P15,6/21/98,N/A,1028723454,Complete,Complete,4672767433,N/A,Complete,Complete,245,N/A,
13,Violet
P16,P16,9/5/01,Complete,2535132609,Complete,Complete,N/A,4725240219,Complete,Complete,
268,82,10,Violet
P17,P17,3/3/77,Complete,4112521449,Complete,Complete,
```

Fig. 4 Sample CSV Code for Hypothetical Patient Data Showing Data Element Structure (different fields separated by commas and different patient records separated by line breaks)

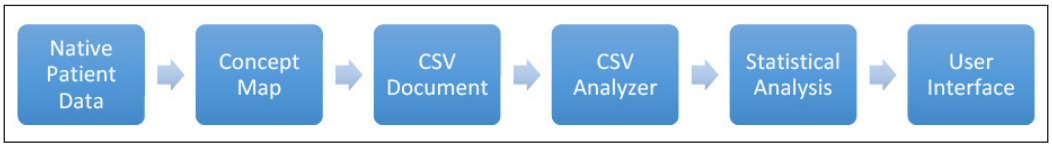

Fig. 5 Flow Chart Representing Data Handling within DCAP

This document was downloaded for personal use only. Unauthorized distribution is strictly prohibited.

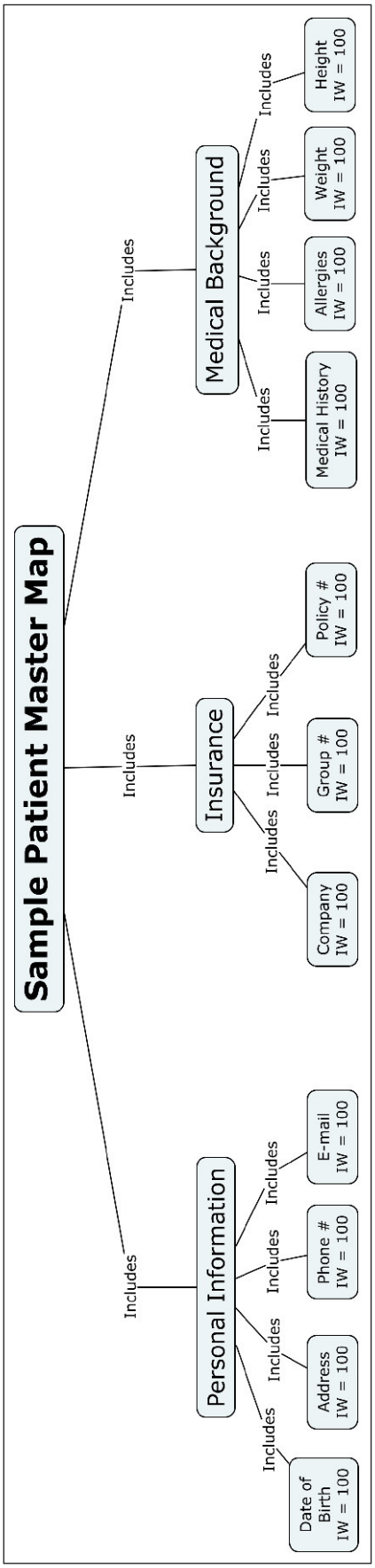

Fig. 6 Sample Patient Master Map with Balanced Scoring (Below Data Field Title)

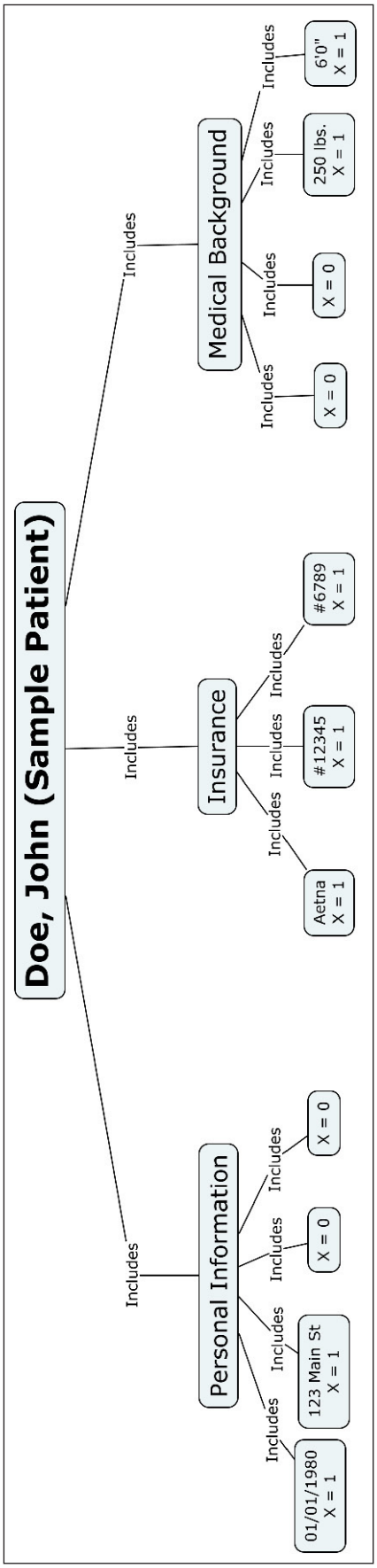

Fig. 7 Sample Patient Map with Balanced Scoring (Labeling Denotes Binary Completeness Variable Below Data Field Title)

This document was downloaded for personal use only. Unauthorized distribution is strictly prohibited.

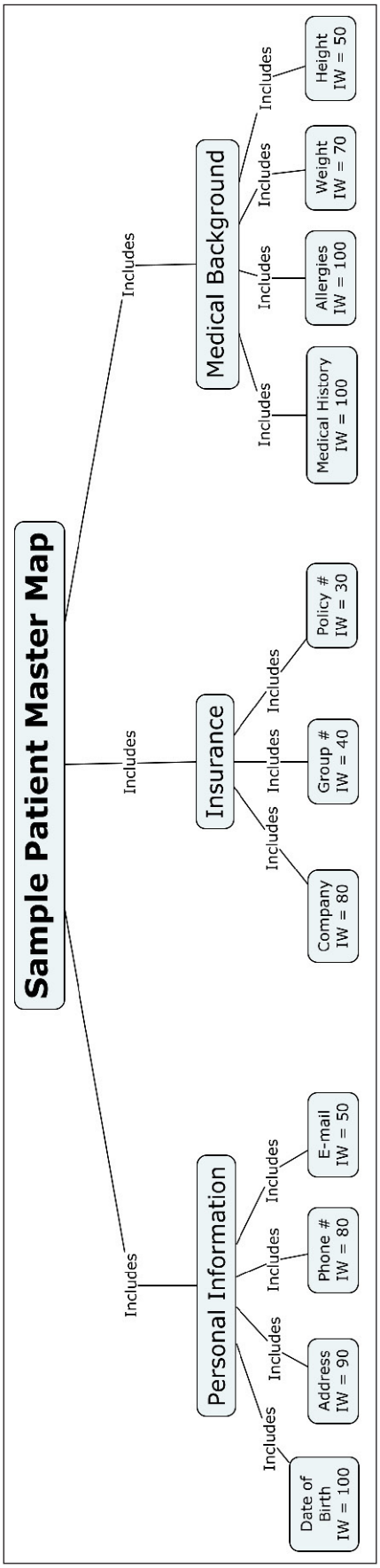

Fig. 8 Sample Patient Master Map with Unbalanced Importance Weights

© Schattauer 2016

This document was downloaded for personal use only. Unauthorized distribution is strictly prohibited.

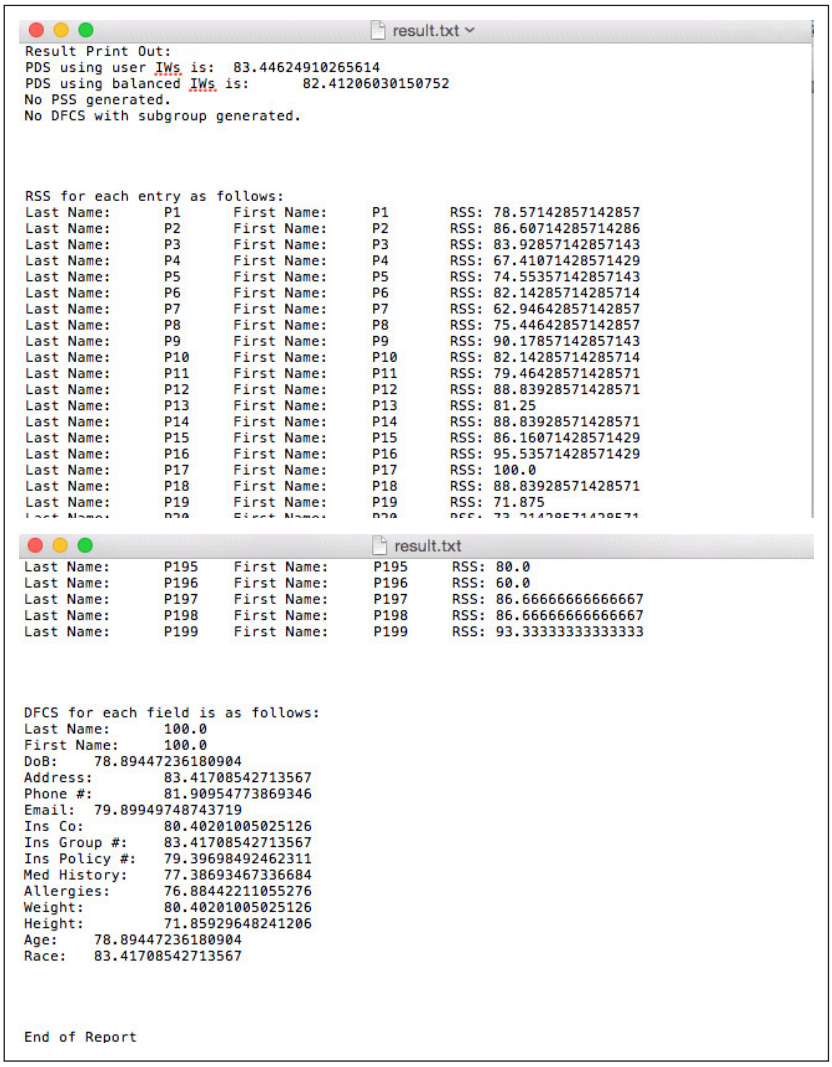

Fig. 9 Sample Hypothetical Patient DCAP Result

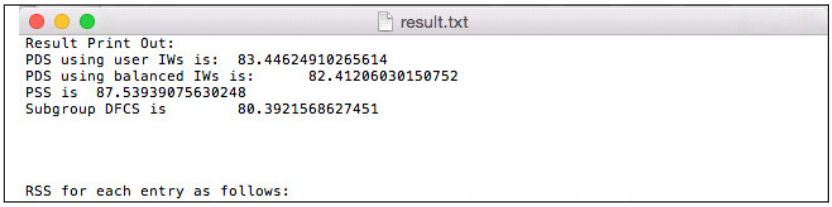

Fig. 10 Sample Patient Result from Hypothetical Data with PSS and Subgroup DFCS scores

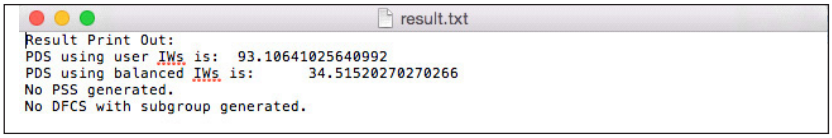

Fig. 11 Balanced IWs for HCUP data results generated

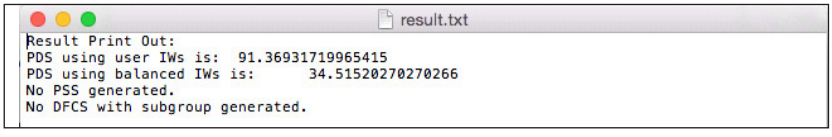

Fig. 12 Unbalanced IWs for HCUP data – results generated

This document was downloaded for personal use only. Unauthorized distribution is strictly prohibited.

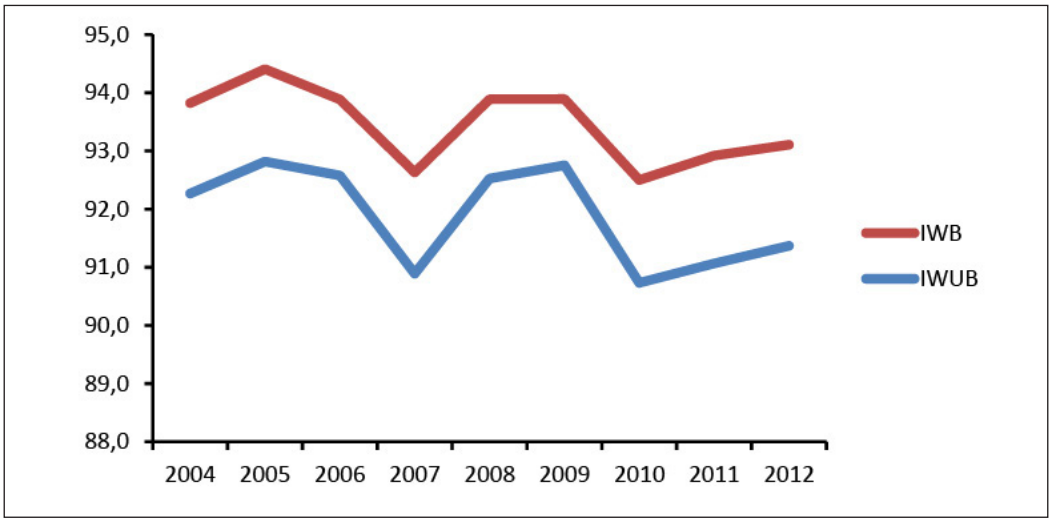

Fig. 13 PDS for HCUP SID Samples from 2004 to 2012

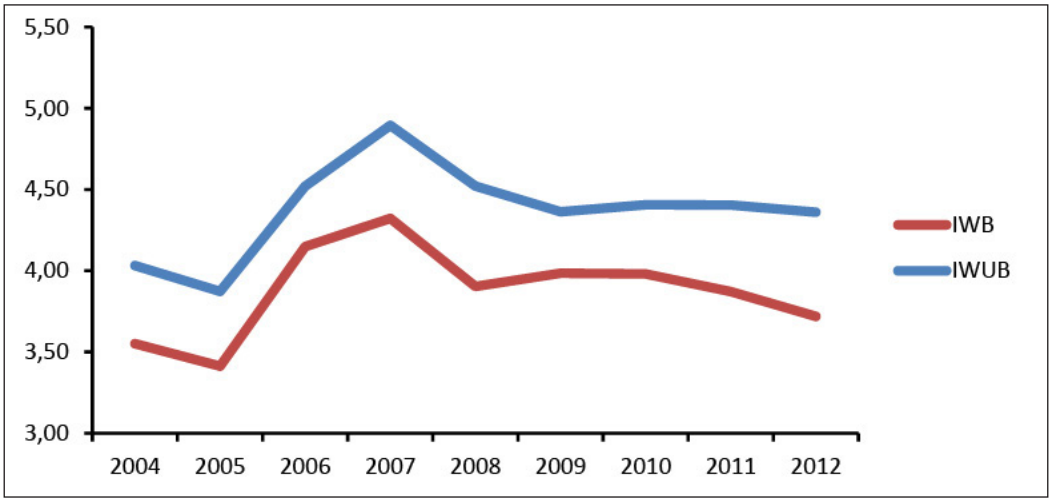

Fig. 14 PDS Standard Deviation for HCUP SID Samples from 2004 to 2012

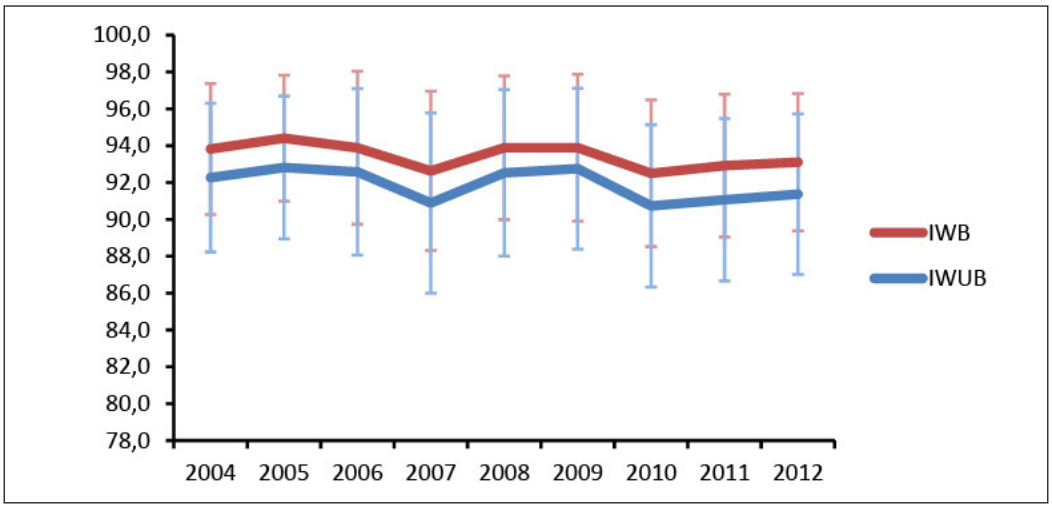

Fig. 15 PDS for HCUP SID Samples from 2004 to 2012 with Standard Deviation errors.

This document was downloaded for personal use only. Unauthorized distribution is strictly prohibited.

Table 1 Comparison of Project with Existing Approaches

Project	Description	Characteristics
Data Completeness Analysis Package (DCAP)	Use concept mapping and advanced statistical analysis to measure data completeness as well as compatibility with various health care database software packages	Only automated to check for completeness, currently exists as a "proof-of-concept" exercise
PCMH Benchmark[1]	Use basic ratios of completeness (patients with given record completed/total patients) to ensure thoroughness	Only checks for completeness, no differentiation on importance of data sets, not automated nor uniform
Private Industry[2]	Variety of solutions to data completeness and accuracy	Not uniform, costly, and often proprietary methodologies used

[1]Benchmark information [28], holistic information [29, 30]
[2]Sources include analytical suites within EMR clinical software (ex. eClinicalWorks) [33], third party vendors (Q-Centrix Data Accuracy Program) [32], and other private sector entities (American Health Information Management Association) [31].

Table 2 Unbalanced Importance Weights for Sample Patient

Data Field	Importance Weight
Date of Birth	100
Address	90
Phone #	80
Email	50
Insurance Company	80
Insurance Group #	40
Insurance Policy #	30
Medical History	100
Allergies	100
Weight	70
Height	50

Table 3 Summary of Various Scores Produced By DCAP

Metric	Definition	Description	Example
Record Strength Score (RSS)	See Equation 1	Measures the strength of an individual patients record	Mr. John Doe's record has a 70% RSS score
Patient Database Score (PDS)	See Equation 2	Measures the strength of the entire patient record database at a health care center	XYZ Medical Center's PDS Score is 50%
Patient Subgroup Score (PSS)	See Equation 3	Measures the strength of the patient records of a specific population of the patients seen by the health care center	The PSS Score of male patients over the age of 55 is 95%
Data Field Completeness Score (DFCS)	See Equation 4 for database See Equation 5 for subpopulation	Measures the strength of recording for a specific data field for either all patients in a database or a subset of patients	The DFCS for Insurance Policy Number is 80% The DFCS for height among females is 86%

This document was downloaded for personal use only. Unauthorized distribution is strictly prohibited.

References

1. Deeks J. Meta-Analysis, Decision Analysis And Cost-Effectiveness Analysis: Methods For Quantitative Synthesis In Medicine. Statist. Med. Statistics in Medicine 1996; 15(14): 1601–1602.
2. Ozcan Y. Quantitative Methods in Health Care Management: Techniques and Applications. Jossey-Bass Public Health. (n.d.). 36.
3. Armitage P. Quantitative Methods In Biological And Medical Sciences: A Historical Essay. Statist. Med. Statistics in Medicine 1996; 15(5): 562–563.
4. Stuart M. Public Health Issues in the Development of Centralized Health Care Databases. Journal of Public Health Management and Practice 1995; 1(1): 57–62.
5. Strauss AT, Martinez DA, Garcia-Arce A, Taylor S, Mateja C, Fabri PJ, Zayas-Castro JL. A user needs assessment to inform health information exchange design and implementation. BMC medical informatics and decision making 2015; 15(1): 1.
6. Martinez DA, Mora E, Gemmani M, Zayas-Castro J. Uncovering Hospitalists' Information Needs from Outside Healthcare Facilities in the Context of Health Information Exchange Using Association Rule Learning. Applied Clinical Informatics 2015; 6(4), 684–697.
7. Ucf-rec.org. UCF Regional Extension Center | Patient Centered Medical Home (PCMH). 2015 [cited 5 March 2015]. Available from: http://ucf-rec.org/services/pcmh/
8. Solimeo SL, Hein MPM, Ono S, Lampman M, Stewart GL. Medical homes require more than an EMR and aligned incentives. The American journal of managed care 2013; 19(2): 132–140.
9. Jha AK, DesRoches CM, Campbell EG, Donelan K, Rao SR, Ferris TG, Shields A, Rosenbaum S, Blumenthal D. Use of electronic health records in U.S. hospitals. New England Journal of Medicine 2009; 360(16): 1628–1638.
10. Walsh S. The clinician's perspective on electronic health records and how they can affect patient care. BMJ 2004; 328(7449): 1184–1187.
11. Shortliffe EH. The evolution of health-care records in the era of the Internet. Medinfo 1998; 98: 8–14.
12. vonKoss Krowchuk H, Moore ML, Richardson L. Using health care records as sources of data for research. Journal of Nursing Measurement 1995; 3(1): 3–12.
13. Hogan W, Wagner M. Accuracy of Data in Computer-based Patient Records. Journal of the American Medical Informatics Association 1997; 4(5): 342–355.
14. Dambro MR, Weiss BD. Assessing the quality of data entry in a computerized medical records system. J Med Syst 1988; 12: 181-187.
15. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: Enabling reuse for clinical research. Journal of the American Medical Informatics Association 2012; 20(1), 144–151.
16. Majid A, Car J, Sheikh A. Accuracy and completeness of electronic patient records in primary care: Family Practice 2008; 25(4): 213–214.
17. Barlow L, Westergren K, Holmberg L, Talbäck M. The completeness of the Swedish Cancer Register – a sample survey for year 1998. Acta Oncologica 2009; 48(1): 27–33.
18. Warsi AA, White S, McCulloch P. Completeness of data entry in three cancer surgery databases. European Journal of Surgical Oncology 2002; 28(8): 850–856.
19. Roos L, Roos N, Cageorge S, Nicol J. How Good Are the Data? Medical Care (n.d.). 266–276.
20. Hassey A. A survey of validity and utility of electronic patient records in a general practice. BMJ 2001; 322(7299): 1401–1405.
21. Arts DGT, De Keizer NF, Scheffer GJ. Defining and improving data quality in medical registries: A literature review, case study, and generic framework. Journal of the American Medical Informatics Association 2009; 9(6): 600–611.
22. Munro B. Statistical methods for health care research. Philadelphia: Lippincott Williams & Wilkins; 2005.
23. Armitage P., & Mainland, D. Elementary Medical Statistics. Biometrika (n.d.). 281–281.
24. Lloyd S. Physician and Coding Errors in Patient Records. JAMA 1985; 254(10): 1330.
25. Critchfield G. Data Entry for Computer-based Patient Records. Aspects of the Computer-based Patient Record (n.d.). 140–145.
26. Prather JC, Lobach DF, Goodwin LK, Hales JW, Hage ML, Hammond WE. Medical data mining: knowledge discovery in a clinical data warehouse. Proceedings of the AMIA Annual Fall Symposium 1997: 101–105.
27. Peute LW, Driest KF, Marcilly R, Bras Da Costa S, Beuscart-Zephir MC, Jaspers MW. A Framework for reporting on human factor/usability studies of health information technologies. Studies on Health Technologies and Informatics, IOS Press 2013; 194: 54–60.

This document was downloaded for personal use only. Unauthorized distribution is strictly prohibited.

28. Recommended Core Measures for Evaluating the Patient-Centered Medical Home: Cost, Utilization, and Clinical Quality Web. 6. 2015 [cited 5 March 2015]. Available from: http://www.commonwealthfund.org/~/media/Files/Publications/Data Brief/2012/1601_Rosenthal_recommended_core_measures_PCMH_v2.pdf
29. Fernald DH., Deaner N., O'Neill C, Jortberg BT, deGruy III FV, Perry DW. Overcoming early barriers to PCMH practice improvement in family medicine residencies. Family Medicine-Kansas City 2011; 43(7): 503.
30. Patient-centered medical home: building evidence and momentum: a compilation of PCMH pilot and demonstration projects. Patient-Centered Primary Care Collaborative, 2008.
31. Ensuring Data Integrity in Health Information Exchange. 2015 [cited 5 March 2015]. Available from: http://library.ahima.org/xpedio/groups/public/documents/ahima/bok1_049675.pdf
32. Q-centrix.com. Patient Data Accuracy Programs for the Healthcare Industry – Q-Centrix. 2015 [cited 5 March 2015]. Available from: http://www.q-centrix.com/why-q-centrix/our-promise/accuracy
33. eClinicalWorks. Population Health (CCMR) – eClinicalWorks. 2015 [cited 5 March 2015]. Available from: https://www.eclinicalworks.com/products-services/population-health-ccmr/
34. Lanzola G, Parimbelli E, Micieli G, Cavallini A, Quaglini S. Data Quality and Completeness in a Web Stroke Registry as the Basis for Data and Process Mining. Journal of Healthcare Engineering 2014; 5(2): 163–184.
35. Dabbagh N. Concept Mapping as a Mindtool for Critical Thinking, Journal of Computing in Teacher Education 2001; 17 (2): 16–23.
36. Jain GP, Gurupur V, Schroeder JL, Faulkenberry ED. Artificial Intelligence-Based Student Learning Evaluation: A Concept Map-Based Approach for Analyzing a Student's Understanding of a Topic, IEEE Transactions on Learning Technologies, 2014: DOI: 10.1109/TLT.2014.2330297.
37. Gurupur V, Jain GP, Rudraraju R. Evaluating Student Learning Using Concept Maps and Markov Chains: Expert Systems with Applications 2015; 42: 3306–3314.
38. Ahrq.gov. Healthcare Cost and Utilization Project (HCUP) | Agency for Healthcare Research & Quality. 2015 [cited 5 March 2015]. Available from: http://www.ahrq.gov/research/data/hcup/
39. SID Database Documentation. Retrieved June 14, 2015, from https://www.hcup-us.ahrq.gov/db/state/siddbdocumentation.jsp
40. Two-Sample T-Test. (n.d.). Retrieved February 17, 2016, from http://www.ncss.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Two-Sample_T-Test.pdf
41. StatPac User's Guide. (n.d.). Retrieved February 17, 2016, from https://statpac.com/manual/index.htm?turl=compareasamplemeantoapopulationmean.htm
42. Gurupur V, Kamdi AS, Tuncer T, Tanik MM, Tanju MN. Enhancing Medical Research Efficiency Using Concept Maps, Editor: Arabnia HR. Advances in Experimental Medicine and Biology, Springer 2011; 696(Part 7): 581–588.
43. MetaMap – A Tool For Recognizing UMLS Concepts in Text. (n.d.). Retrieved March 06, 2016, from https://metamap.nlm.nih.gov/
44. Nasir A, Gurupur V, Liu X, Qureshi X. Managing Healthcare Patient Data Using the Data Completeness Analysis Package (DCAP), Proceedings of SDPS 2015 Conference, November 1– 5, 2015, Fort Worth, TX.

This document was downloaded for personal use only. Unauthorized distribution is strictly prohibited.