A REVIEW OF CURRENT LITERATURE THAT MEASURES AGREEMENT BETWEEN PATIENT SELF-REPORT DATA AND A REFERENCE STANDARD

By

Thad P. Jarmon

A BMI581 Capstone Paper

Presented to the
Department of Medical Informatics and Clinical Epidemiology and the
Oregon Health & Science University School of Medicine
in partial fulfillment of the requirements for the
degree of Master of Science

March 2020

School of Medicine

Oregon Health & Science University

CERTIFICATE OF APPROVAL

_____

This is to certify that the Capstone paper of

Thad P. Jarmon

has been approved

Mentor/Advisor_____

Table of Contents

ABSTRACT

Objective

The purpose of this paper is to review current literature that measures agreement between patient self-report data and a reference standard and to analyze the similarities and disparities between the results of these studies.

Methods

A clinical literature review of articles measuring agreement between patient self-report and reference standard data for the presence of a clinical diagnosis was performed. Data was extracted from 41 published journal articles containing agreement values for 206 condition-level analyses (cases) across 57 distinct conditions. Cohen's kappa was calculated for each of the 206 cases. Individual cases were sorted into logical grouping such as diagnosis, self-report type and reference standard type. Kappa values were compared within groups to determine which, if any, factors had the greatest effect on agreement.

Results

Several factors across the subgroups we analyzed were had statistically significant differences within their respective groups. Among diagnoses groups, diabetes had a positive effect on agreement. Reference standard data from patient registries also had a positive effect on agreement while biometric data take at the point of service had a negative effect on agreement. No statistically significant differences were found for different types of patient self-report methods or geographic location.

Conclusion

Patient self-report and reference standard data sources each have strengths and limitations, however, the lack of completeness in often found in reference data sets makes it challenging to measure the true level of agreement between these sources. Based on our research, however, it is clear that patient self-report data is at times more accurate than reference standard data and can serve as an effective means of augmenting reference standard when collected and captured effectively.

INTRODUCTION

In recent years, electronic methods for capturing and maintaining clinical data have become the standard for healthcare providers.[1] Consequently, clinical databases have become larger, more easily accessible, and more frequently used to measure the performance of healthcare providers.[2] Creating a complete and correct view of patient history by using a combination of data sources is important for clinical, financial and regulatory reasons. Incomplete or inaccurate data may cause providers to, for example, miss a crucial co-morbidity or underlying risk factor, or report inaccurate data to insurers. Despite the growing need for patient information, there are very few databases that provide complete coverage of patient medical history over more than a few years.[3] Most databases that are readily available for research and clinical decision making are open systems that only contain either medical history occurring within a limited timeframe or within a respective organization's facilities.[4] Fully-realized health information exchange would be the ideal solution to this problem, but until the majority of providers and insurers agree to enforce data sharing standards and eliminate data blocking, this scenario is unlikely.[56] For organizations using smaller, incomplete databases, augmenting medical history with patient self-reported clinical data may be a way to close the information gap.[7]

While patient self-report may have the potential to augment medical records, there are also limitations to self-reported medical data.[89] Specifically, previous research has suggested that self-reported data is not accurate enough to supplement medical records and may, therefore, add confusion to a patient's medical history.[101112] Among concerns detractors have about the reliability of patient self-report data are patient recall bias, social concerns that may cause patients to misreport data, and the average level of health literacy and engagement of the patient population.[131415] Before patient self-report data can be applied *en masse* to fill in the blanks of existing data sets, it is important to identify how well this type of data corresponds to data already contained in existing clinical information, referred to as a reference standard, and what factors increase or decrease accuracy. To better understand the quality of patient self-reported data, multiple researchers have measured agreement between self-report data and reference standard data and attempted to explain why the two sources often diverge. These studies often look at agreement for medication, disease or illness, or lifestyle and behavior.

Despite the growing body of work devoted to understanding agreement between patient self-report and reference standard data, there are few articles which provide a summary of the results of these articles in aggregate. To this end, **the purpose of this paper is to review current literature that measures agreement between patient self-report and a reference standard and to analyze the similarities and disparities between the results of these studies**. Providing this type of analysis may be useful to future research as well as to help inform readers how best to develop interview questionnaires and to determine what types of reference standard data will yield the most accurate results.

METHODS

To measure concordance between illnesses captured in a reference standard and self-reported data gathered from patients, we conducted a systematic literature review and initial analyses of journal articles published within the past fifteen years. The objective of this analysis was to determine whether there are patterns that suggest certain characteristics in either a patient self-report instrument or a reference standard that may lead to higher agreement between the two sources. To discover possible patterns, we analyzed the relationships between variables including diagnosis, reference data type, self-reported data type, and features of the study populations.

Literature search and review

To identify papers for inclusion in the research, we searched the MEDLINE database to find papers using relevant MeSH search terms. Given the improvement in EHR technology and the rapid increase in EHR utilization, we limited our results to research conducted within the past fifteen years. In general, results from research conducted earlier than 2000 may not be applicable to the realities of today's patient and provider population. The inclusion and exclusion criteria as well as search terms that were used are listed below:

Inclusion criteria

1) Peer reviewed, English language articles published since 2000.

2) Articles must have compared patient self-reported data (via a survey instrument) to a recorded medical condition in the form of a confirmed diagnosis (documented with an ICD9 or ICD10 code, having been noted in a clinical record or medical chart, or identified using biometric data values).
3) Patient populations were reported for all measured diagnoses.
4) All confusion matrix values (TP/TN/FP/FN) were provided (or enough data was available through a combination of partial matrix data and reported agreement metrics to derive all matrix values).
5) At least one agreement metric for each measured diagnosis (kappa, sensitivity, specificity, PPV or NPV) was published by the authors of the study.

Exclusion criteria

1) Articles that compared self-report to behavior, drug use or diagnostic services rather than to a medical diagnosis were excluded (articles comparing self-report data with a mental health diagnosis were also excluded due to the complexity of accurately diagnosing patients).
2) Articles that compared agreement for pediatric patients were excluded as parents often provided self-report data rather than the patients themselves.
3) The reference data used for research consisted of data from a time period earlier than January 2000.

Medline search and results (as of November 2018)

(("self report*") AND medical record*) AND (agreement or validation or validity) AND (kappa or PPV) AND ("2000/01/01"[PDAT] : "3000"[PDAT])

Our initial search returned 1,738 matches, resulting in high recall but very low precision. In order to exclude entries that were only meeting notes or study abstracts rather than full journal articles, we excluded articles that were not also found through PubMed®. Next, to remove irrelevant literature from our pool of candidates, the research team reviewed the titles and abstracts for content which clearly did not meet our needs. Because our research was designed to look at agreement between the presence of a disease or illness, we reviewed the papers for concepts within the title or abstract, such as medication or behavioral terms like

smoking. If we found keywords that suggested the article might not meet our inclusion requirements, we scanned the article to determine if the article should be excluded. If no potentially excluding keywords were found, we allowed the article to remain in the pool for further review. Next, we ensured the article compared patient self-reported diagnosis data to a reference standard that also reported the presence of a patient diagnosis. Once articles without relevant subject matter were eliminated, the team further reviewed the articles for raw data values and/or agreement metrics which would allow us to derive a confusion matrix (or error matrix) resulting in a final total of 41 articles containing 193 condition-level analyses for 50 distinct conditions. Shown below is a flow diagram outlining our search, evaluation and inclusion process.

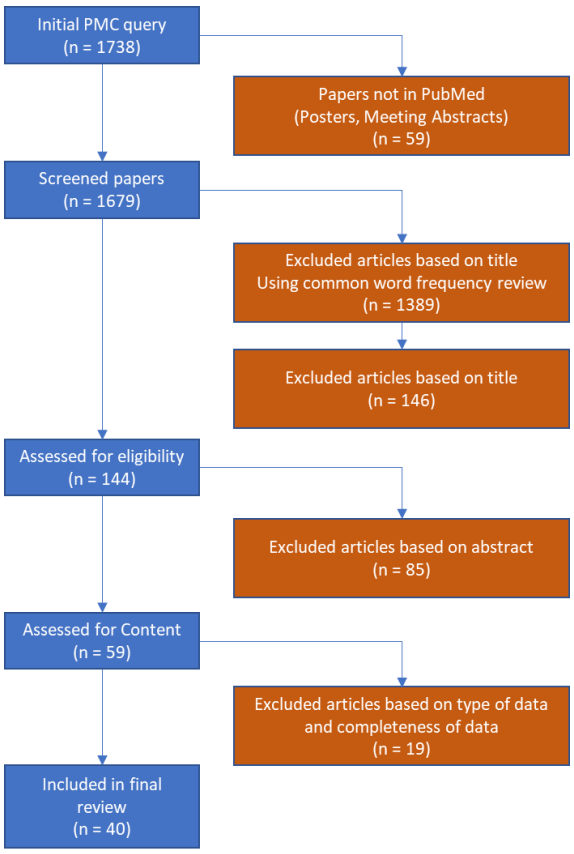


Figure 1. Flow diagram for article inclusion

Extraction and curation of agreement data

Next, the research team extracted the values to be used to calculate agreement, including the study population (n), true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). Agreement metrics including positive predictive value (PPV), negative predictive value (NPV), sensitivity and specificity, total agreement, and Cohen's kappa were also extracted. When the raw input values (TP/TN/FP/FN)) were not provided in an article, the team derived the input values from agreement metrics using Microsoft Excel solver when possible. Studies where the raw input values could not be calculated were excluded from further analysis. Additionally, articles that did not provide a kappa value were excluded. Once all confusion matrix values were extracted, the team recalculated sensitivity, specificity, kappa, and standard error to check the consistency of results, thereby ensuring that cases used included the same set of measurements and that a single method was used to calculate the measurements. The confusion matrix values were then back-tested to ensure that agreement metrics could be replicated to a reasonable degree of accuracy.

Cohen's kappa was calculated as follows:

$$\kappa \equiv \frac{p_o - p_e}{1 - p_e} = 1 - \frac{1 - p_o}{1 - p_e},$$

where (Po) represents the actual observed agreement, and (Pe) represents chance agreement. The standard deviation of kappa was calculated as: $\kappa - 1.96 \times SE\kappa$ to $\kappa + 1.96 \times SE\kappa$, where SE = standard error.

Although the use of kappa as a measure of agreement has been criticized, it is nonetheless a generally agreed upon measure for concordance and will be the measure used for reporting results within this paper. The value ranges used to classify the relative strength of kappa can also differ between studies. Earlier studies generally use a system proposed by Landis and Koch while more recent studies may diverge from that standard and use a different interpretation, such as that proposed by McHugh, which provides greater detail and uses more contemporary terminology for classification.[16][17] In this paper, the McHugh-based kappa ranges will be used, as follows: $\kappa = 0.0–0.20$: no agreement, $\kappa = 0.21$-0.39: minimal agreement, $\kappa = 0.40–0.59$: weak agreement, $\kappa = 0.60–0.79$: moderate agreement, $\kappa = 0.80–0.90$: strong agreement: $\kappa > 0.90$: almost perfect agreement.

Feature extraction and dimensionality reduction

To facilitate our analysis, we grouped reference standard and self-reported data into various strata based on logical subgroups. Cases were stratified by self-report type, reference standard type, geographic region, high-level diagnosis groups, study population type and time frame. Additionally, given the number of disparate cases, creating groupings helped to increase statistical power of our results by increasing the size of the sample within a given group.

Diagnosis groups

While some studies provided a list of ICD9 or ICD10 codes that defined the set of diagnoses that comprised a measured condition within a reference standard, in most cases, no clear set of rules was given to identify the parameter of a specific condition. This was almost always the case with medical conditions measured in patient self-report data. Therefore, to facilitate analysis by diagnoses, we sorted each diagnosis into a logical category when possible, such as grouping hyperlipidemia and high cholesterol under a single group.

Self-report data types

Self-report data were grouped as either **self-administered** or **interviewer administered**. In both cases, self-reported data is captured in a survey instrument - usually in the form of a questionnaire that is either completed by the patient without any outside assistance (self-administered) or administered in a face-to-face interview or via telephone by a trained interviewer (interviewer administered). One challenge we found in differentiating patient self-report methods below a single stratum was determining the level of interaction between the patient and the interviewer or clinical staff. Even for studies with trained interviewers, little information is given about how much assistance an interviewer provided to respondents. Similarly, little information is given about whether patients taking self-administered surveys received any assistance on understanding survey questions or entering survey responses. Thus, determining at an individual case level how autonomous patients were when self-reporting was nearly impossible.

Reference standard data sources

In contrast, reference standard data is frequently taken from legacy clinical or administrative healthcare data reporting systems such as electronic- or paper-based medical records. As with patient self-report, we

also grouped reference standard data into sub-categories based on the underlying data source: medical records, administrative data, registry data, or biometric data. Clinical care data extracted from either paper-based charts or electronic systems designed specifically to capture patient clinical information was classified as **medical records** data. Clinical data that were extracted from systems that capture diagnosis and procedure codes as a means for admissions, reimbursement or purposes other than patient care are grouped as **administrative** data. **registry** data refers to data extracted from databases created from a patient registry. Registry data were differentiated from general medical records data because registries are designed for analytical use and will consequently have a higher level of detail, accuracy and completeness than a typical medical record. Finally, data from clinical tests such as blood pressure, glucose levels, or cholesterol levels that were obtained with the specific intent of being used as a reference standard data were classified as **biometric** data.

Other groupings

Several other groupings were employed as well to help ascertain what factors might improve agreement. To determine whether social or technological differences in agreement occurred based on regional distinctions, cases were grouped by major geographic regions including: North America, Europe and Asia-Pacific. Finally, cases were also grouped based on whether agreement was being measured for the general population or within a patient population that was required to have been previously diagnosed with a specific condition. Furthermore, if the condition being measured was related to the study population's underlying condition, the group was then also grouped into a separate category. As an example, agreement for self-reported hypertension in patients diagnosed with congestive heart failure (CHF) might possibly be greater than agreement in the general population because of increased health awareness within the CHF group.

Statistical Analyses

To determine if any or all subgroups individually played a significant factor in predicting the level of agreement between patient self-report data and reference standard data, we performed a univariate generalized linear model (GLM) procedure in SPSS to determine the impact on kappa of the following factors: self-report type, reference standard type, diagnosis group and region (a discussion of subgroupings

can be found in the following section).[18] These were performed on the entire set of studies with the exception of diagnosis group, which was limited to diabetes, hypertension, hyperlipidemia, stroke and myocardial infarction because of low sample volume in other diagnosis groups.

RESULTS

We collected data from 42 studies measuring agreement for variety of conditions including hypertension, diabetes, high cholesterol, myocardial infarction among others. Across articles that were selected for review, there were of 193 instances (cases) of agreement measured for 50 unique diagnoses. Of the 50 diagnoses measured, five diagnoses (diabetes, hypertension, hyperlipidemia, stroke and myocardial infarction) made up almost half (47.7%) of the cases. Because many of the remaining diagnoses were either non-specific (such as kidney disorder) or only included one or two cases, the main focus of the results and discussion will be limited to these five diagnoses unless explicitly stated.

Of the 193 unique cases analyzed, 114 cases (59.1%) were from the U.S. and Canada, 40 (20.7%) were from Asia and Australia, 29 (15.0%) were from Europe, and the remaining 10 (5.2%) were from other regions. Sample sizes varied greatly across studies, with the largest study consisting of over 12 million patients and the smallest study consisting of only 57 patients.[19][20] In general, the samples in studies conducted in the United States were among the smallest, averaging around 2,000 patients. Studies conducted in European countries averaged slightly higher with roughly 3,000 patients, whereas, studies conducted in Canada, Asian countries, and Australia had significantly higher patient populations averaging between 20,000 and 25,000 patients (after excluding the South Korean study that included over 8 million patients).

With regards to patient self-report type, self-administered surveys (paper-based or mail-in surveys) comprised more than half of the cases (52.8%), and 45.6% of cases were face-to-face or telephone interviews. Medical records (electronic or paper-based) were used over half the time as reference standard (57.5%). Insurance and administrative claims were used 31.6% of the time. Other sources of reference data included biometric data (8.3%), chart abstraction (4.7%) and registry data (2.6%). The full descriptive statistics can be found in the appendices.

Comparison of derived kappa and reported kappa

In general, the differences between the published kappa, as reported in the original articles, and our re-calculated kappa were negligible. Using the formula for Cohen's kappa noted in the methods section resulted in over 90% of cases with less than two percent variance between our results and the published results; therefore, we feel confident that our kappa calculation provides an accurate method of normalizing results from the data set.

Table 1. Comparison of published kappas and derived kappas.

% Variance	Cases	% of cases
0%	29	15.6%
<=1%	113	60.8%
<= 2%	26	14.0%
<=5%	11	5.9%
>5%	7	3.8%
Total	**186**	**100.0%**

Note: In nine cases a kappa was not published in the article; therefore, a comparison could not be made.

Level of agreement by diagnosis

Of the five main diagnoses found in our pool of studies, diabetes had the highest level of agreement amongst the cases we reviewed. The average kappa for diabetes was strong (κ =0.80). Myocardial infarction, hypertension and stroke all had weak agreement at k =0.56, κ =0.53 and κ =0.51 respectively. Hyperlipidemia had the lowest level of agreement amongst the five conditions with an average kappa of 0.34. Univariate analysis indicated there was a significant difference in kappas between the five main diagnoses. Diabetes was found to have a significant effect on the level of agreement ($p<0.05$) (see Appendix A).

Table 2. Average kappa for each of the five main conditions.

Diagnosis	Agreement	Avg. kappa	Std. dev. of kappa	Number of cases
Diabetes	Strong	0.80	0.10	29
Myocardial infarction	Weak	0.56	0.18	12
Hypertension	Weak	0.53	0.21	30
Stroke	Weak	0.51	0.15	13
Hyperlipidemia	Minimal	0.34	0.24	13

Level of agreement by geographic region

When measuring agreement as a function of geographic location across all conditions, kappa in Europe and North America are nearly equal (κ =0.54 and κ =0.53, respectively) and Asia Pacific countries have a combined kappa of 0.45. When only considering the five core conditions, kappa in both European, North American and Asia Pacific studies are slightly higher (κ =0.58, κ =0.64 and 0.52 respectively). No region was found to have a significant effect on the level of agreement (Appendix A).

Table 2. Average kappa by region for core diagnoses compared to overall kappa by region (Dx = diagnosis).

	Core Conditions						Total for All Dx
Region	**Diabetes**	**Hyper-lipidemia**	**Hyper-tension**	**Myocardial infarction**	**Stroke**	**Total for Core Dx**	
Asia Pacific	0.71	0.26	0.55	0.46	0.41	0.52	0.45
Europe	0.88	0.09	0.60	0.79	0.42	0.58	0.54
North America	0.82	0.43	0.64	0.51	0.52	0.64	0.53
Other	0.93	0.46	0.27	-	-	0.37	0.40

Level of agreement by self-report type

Self-administered surveys (κ = 0.52) generated higher agreement than face-to-face interviews (κ = 0.49) for the five core conditions. Additionally, self-administered survey kappa was higher in all conditions except MI, where agreement was almost similar (MI self-administered survey κ = 0.55, MI face-to-face κ = 0.57). Neither self-administered surveys or face-to-face interviews were found to have a significant effect on the level of agreement (Appendix A).

Level of agreement by reference standard type

Medical records (including EHR and chart abstraction) had higher agreement (combined κ = 0.64) than both administrative data which includes health insurance claims, Medicare claims and hospital admissions data (combined κ = 0.47) and biometric data which includes blood glucose and blood pressure readings (κ

= 0.40). Both registry data and biometric data were found to have a significant effect on the level of agreement (Appendix A). Registry data had a positive level of agreement while biometric data had a negative effect on the level of agreement.

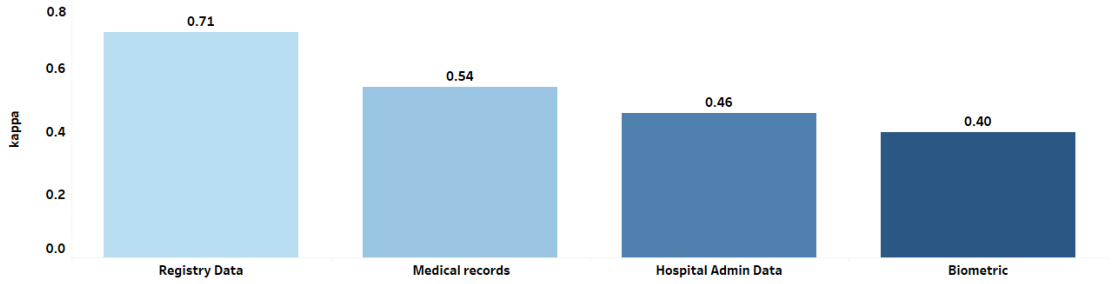


Figure 2. Level of agreement by reference standard type**.**

Level of agreement by general health of the study population

The mean agreement for studies that measured the general population was κ =0.48 and only slightly higher when the study population consisted of patients that were all verified as having a similar condition, κ =0.54. Even when the measured diagnosis was related to the patient's underlying condition, agreement was still similar (κ =0.55)

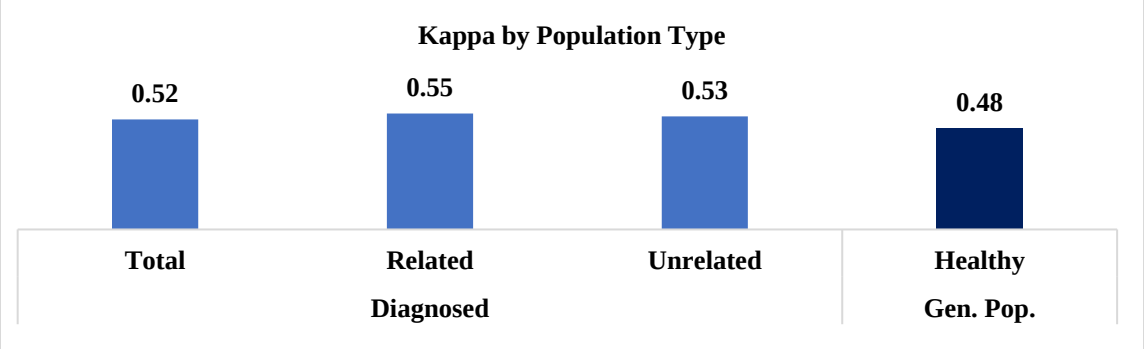


Figure 3. Level of agreement by general health of the study population**.**

DISCUSSION

In reviewing recent journal articles that measure agreement between patient self-report and a reference standard, our goal was determine what role specific factors such as diagnosis, region, data type or survey

instrument played in the level of agreement between the two sources. To this end, we conducted a systematic literature review and analyses of data reported in papers published within the past fifteen years.

The two factors that appear to have the greatest effect on agreement are the condition measured and the type of reference standard data used. Diabetes was the only condition with strong agreement between self-reported data and a reference standard. This may be due to the more chronic and symptomatic nature of diabetes and the dosing frequency of diabetes therapy when compared to the other conditions we measured. In terms of reference standards, chart abstraction and electronic medical records had the highest level of agreement with patient self-report data. Medical claims data sources such as Medicare claims, hospital administrative data and insurance claims data had much lower agreement while biometric data had the lowest level of agreement. These finding are understandable considering the level of detail found in charts and structured medical records when compared to administrative data sources.

Challenges of measuring agreement with self-report

Two challenges that occur when conducting research on agreement are determining what data sources to use and how best to measure agreement. Across the studies that were analyzed as a part of our research, the most ubiquitous measure of agreement was Cohen's kappa (κ), a measure of inter-rater reliability.[21] Cohen's kappa provides researchers with a way to employ a single measure to assess concordance between sources and allows a simple way to compare the results from one study to other studies. While there are valid criticisms about using kappa to measure agreement when neither source is certain to be accurate, we nonetheless chose kappa as our means of comparison for three primary reasons[22]. Firstly, Cohen's kappa is widely used and generally accepted by researchers in this area as a suitable measure of agreement. Secondly, using kappa allowed us to validate our data extraction and calculation methods by comparing our results to those of the published articles. Thirdly, once we were satisfied that our kappa calculations were correct, we could identify articles in which kappa results may have been misprinted or mislabeled, thereby serving as a check on the published results.

Possibly of even greater importance than determining how to best report agreement may be the challenge researchers face in identifying a standard methodology for capturing and collecting data. This challenge occurs in both implementing a best-practice for collecting self-report data and in selecting the best

reference standard data source to use for comparison. For example, the methods for collecting patient self-report information vary from a simple self-completed paper survey that was filled out by patients in a waiting room to more sophisticated research instruments administered by trained interviewers in a face-to-face setting where ambiguities are resolved immediately.

Data types

Contrary to our initial assumptions, self-administered surveys generated higher agreement than face-to-face interviews for the five core conditions. These findings show that while some patients may be initially confused by a survey question, further clarification provided by an administrator may not improve agreement suggesting that other factors, such as patient recall or the accuracy of medical records, may be more significant in determining overall agreement. Conversely, this finding may also be a result of interviews eliciting even more complete information than self-administered surveys such that the level of false negatives are higher when compared against an incomplete reference standard. A potential confounder to this hypothesis is the likely amount of heterogeneity among methods used to collect self-administered survey data. Often, there was no discussion provided on the deployment of self-administered surveys beyond simply stating that a particular survey was mailed or a paper-based survey was used to collect data. Due to this lack of information, it is difficult to assess the actual level of assistance offered by providers to survey respondents, thus creating a more specified subgroupings within the self-administered survey group was not possible.

Different types of reference standard data, called “gold standard” by some, also has its strengths and limitations. Most commonly, patient diagnoses have been documented by clinical or administrative staff using ICD9 or ICD10 diagnosis codes. While this disease classification system provides a standardized method for documenting illnesses, ICD codes may sometimes be used out of necessity or convenience when a lack of specificity prevents a more precise description, especially in administrative data sources.[23] This issue can potentially lead to a discrepancy between the patient’s actual condition and the documented diagnosis.

Another reference standard data source is data obtained directly from clinical notes, usually manually reviewed and extracted from paper-based records or EHR systems. While clinical notes often provide more

detail than an ICD10 code, variations in terminology may create ambiguity - thus also possibly leading to disagreement.[24]

Finally, the use of biometric data as a reference standard also has its challenges when used to measure agreement. In patients who have never been formally diagnosed with a condition, there is no formal clinical basis for the patient to affirm a particular condition. Therefore, in most cases, the patient would have denied a diagnosis that discover via biometric data - leading to disagreement between self-reported data and the biometric reference. While this is not a criticism of biometric data as a means of documenting a patient's condition, inter-rater reliability will likely decrease due to an increased number of false negative reports.

Another concern with reference standard data is the difficulty in effectively enforcing strict time alignment with patient self-report data.[25] For example, in a case where reference standard data is only available for the past year, patients may self-report conditions that were documented over a decade ago - even if asked only to consider the past year. Unsurprisingly, given the large variety of methods, there is a high level of disparity in agreement between studies because of differences in reference data sources, patient self-report type, disease or illness, patient characteristics, or even country. Although there does appear to be noticeable and statistically significant trends between the level of agreement for specific diseases, such as diabetes and hypertension, even when measuring agreement for same condition among similar patient cohorts, kappa can vary significantly.

Condition

Agreement for diabetes was only condition with strong agreement between self-reported data and a reference standard. This may be due to the more chronic and symptomatic nature of diabetes and the dosing frequency of diabetes therapy when compared to the other conditions we measured. Two other chronic but less symptomatic conditions, hypertension and hyperlipidemia both had much lower agreement with a kappa of 0.34 and 0.53 respectively. Surprisingly, when combined together, agreement for cancer patients was only $\kappa = 0.64$. Although prostate and breast cancer both showed high agreement ($\kappa = 0.78$ and $\kappa = 0.81$ respectively) there was only one study each for these conditions. Lung cancer had the lowest level of agreement with $\kappa = 0.43$. Similarly, acute conditions such as myocardial infarction and stroke had markedly

lower agreement than diabetes as well. The results of our analysis may suggest that a combination of the symptomatic nature of diabetes, the frequency of treatment and lifestyle changes that occur as a part of diabetes management play a role in concordance between patients and reference data.

Geography

Several Asian and European counties have had nationally-based health systems for decades allowing researchers to advantage of nationally deployed health surveys or had access to decades of data captured from national health insurance databases. Conversely, many of the more recent studies done in the US were only able to access limited information from relatively small patient populations using small EHR data sets or medical record repositories which did not capture all of the patient's clinical history. Due to the real-world limitations of acquiring self-report and reference standard data, researchers must currently temper their expectations for agreement between the two data types. When analyzing data reported for the five core conditions by geographic region, it does not appear that geography played a significant role in agreement since similar agreement trends are seen across regions. One notable exception to this is strong agreement for myocardial infarction (MI) in European studies with an average kappa of 0.79. The results, however, are based on data from only two studies.

Population Health Status

Comparing inter-rater reliability results by population type, the level of agreement was relatively similar. It does not appear that patients with a preexisting condition had a higher level of agreement between self-report data and reference data, even when the measured condition was related to the patients existing condition.

Case Study: Causes of disagreement in common conditions: an in-depth review of four similar studies

Even when measuring agreement for the same diagnosis, differences in methods or study population can increase or decrease the level of agreement between patient self-report and reference standards. To better illustrate how the differences between various methodologies can affect agreement, the table shown below contains a comparison of a subset of articles analyzed as a part of our research. The four studies highlighted below measure agreement in hypertension, hyperlipidemia or diabetes and contain a mix of self-report data

from national health surveys that were administered by trained interviewers, self-administered patient surveys, reimbursement claims data that were captured as a part of a national, single-payer healthcare system, EHR data captured within a single clinic, and biometric data captured at the point of service.

Table 3. Country, age and kappa by diagnosis for articles in agreement case study.

				Agreement (κ)		
Article	**Country**	**Age**	**Participants**	**Hypertension**	**Diabetes**	**Hyperlipidemia**
Wu	Taiwan	41.2	15,574	0.69	0.76	0.32
Peterson	Australia	51.9	7,269	0.21	0.58	(0.02)
Tenkorang	China	60.3	13,561	0.26	n/a	n/a
	India	52.1	10,870	0.14	n/a	n/a
	Russia	62.4	4,081	0.45	n/a	n/a
	South Africa	60.4	3,908	0.07	n/a	n/a
	Ghana	60.2	5,069	0.12	n/a	n/a
Malik	US	56.6	230	0.51	0.85	0.48

The first article, a 2014 study by Wu et al., measures agreement between patient self-report and reimbursement claims data from a national health insurance database in Taiwan.[26] The researchers collected data from over 15,000 participants to measure agreement for diagnoses, medication use, and health-system utilization. Fourteen separate diagnoses were measured including hypertension, diabetes, stroke, and high cholesterol. The study found moderate agreement for hypertension and diabetes ($\kappa = 0.69$, $\kappa = .76$), and minimal agreement for hyperlipdemia ($\kappa = 0.32$). In contrast to the Wu study, a 2016 study by Peterson et al. measured the prevalence of risk factors for chronic conditions in Australia.[27] The study analyzed over 7,000 adult patients for cardiovascular disease risk factors, comparing biometric data captured as a part of a national survey to patient self-reported prevalence of hypertension, high cholesterol, and diabetes. Rather than rely on claims data as a reference standard, Peterson chose to use biometric values such as blood pressure to serve as a de facto patient diagnosis. Overall, Peterson found low agreement between self-reported data and biometric data with minimal agreement for hypertension ($\kappa = 0.21$), nonexistent agreement for high cholesterol ($\kappa = -0.02$), and weak agreement for diabetes ($\kappa = 0.58$). In this case, the Australian national database may be somewhat suspect in that the authors question the accuracy and veracity of the data due to inconsistent user input of clinical information into the EHR system. In contrast, Wu felt that Taiwanese patient self-report may be hampered because of cultural reasons. In some cases, a

patient's family might be consulted about a condition rather than the patient and the family might opt to not inform the patient about an existing disease, thereby, increasing the likelihood of a false negative report.

While Wu and Peterson both measured agreement across multiple diagnoses, a 2017 study by Tenkorang et al. measured agreement for a single diagnosis, hypertension, in multiple countries using a consistent method to collect data.[28] The study measured concordance between self-reported diagnosis and biometric data for patients in five countries in Africa and Asia, including South Africa, Ghana, India, Russia and China. The study used self-reported data collected in 2007 and 2008 from patients who participated in the World Health Organization's Study on Global Ageing and Adult Health (SAGE). Although kappa was not calculated by the authors, the article included the proportion of patients who correctly or incorrectly reported having (or not having) a disease; therefore, we were able to calculate kappa from the data provided. Overall, Russian patients also had the highest agreement ($\kappa = 0.45$) followed by China ($\kappa = 0.26$) with India, Ghana and South Africa all having very low agreement ($\kappa = 0.14$, $\kappa = 0.12$ and $\kappa = 0.07$ respectively); results that generally correspond to the relative literacy rates among the five countries. These findings suggest that socioeconomic factors may play a significant role in data agreement levels.

The final study, published in 2011 by Malik et al., compares agreement between self-reported data and EHR data for nineteen comorbidities in patients with heart failure.[29] This study differs from the previous studies in several areas. First, the sample size is much smaller - only 230 patients were included in the study. Secondly, the study population is based on patients that were all diagnosed with a common condition, heart failure. This study had the oldest average age of 56.6 years and measured agreement for comorbidities that were related to heart failure, such as hypertension and myocardial infarction, as well as unrelated comorbidities such as diabetes. Additionally, the patient population is based in the United States and uses reference data from EHR data from a single medical center. Self-report data was taken from a self-administered medical history survey rather than from an interviewer administered national survey Malik found weak agreement for hypertension ($\kappa = 0.5$) and dyslipidemia ($\kappa = 0.48$) and strong agreement for diabetes ($\kappa = 0.85$). The findings align with similar studies that measured agreement in disease-based cohorts in that agreement appears to be higher in these types of patients because they likely have a higher level of personal healthcare awareness.

CONCLUSION AND NEXT STEPS

In summary, the disparity between the results of four articles reviewed suggest that a wide variety of factors can influence agreement between self-report and reference standard data including clinical, demographic, socioeconomic and cultural factors. While self-report type and the quality and currency of reference standard type can play a role in improving agreement, it is not clear that any one factor, or set of factors consistently plays the main role in agreement across all conditions and geographic regions.

As mentioned earlier, patient self-report methods and reference standard data sources each have inherent strengths and limitations. Although patient self-report presents a practical and cost-effective way of augmenting medical records, several researchers have noted that age, patient mistrust, a lack of overall healthcare awareness, unfamiliarity with medical terminology, and issues with recall all play a role in misinformation.[3031] Reference standard data also has its strengths and limitations including currency, accuracy and completeness Whether intentional or not, the general proportion of false negatives, especially among conditions such as cancer and chronic heart failure is surprising and suggests that neither existing means employed by providers to capture clinical data at the point of service nor patient self-report alone is sufficient as a means to document medical history.

Although, in this paper, we only reported descriptive data and conducted univariate analyses, we have enlisted several of our colleagues to perform a full meta-analysis, including multivariate analyses to better understand to interplay between methods, data types, location and patient types and how the combination of these entities affects agreement. While the results of the meta-analysis will not be completed in time for inclusion in this paper, we anticipate the results will be submitted for future publication and hope the results will contribute to the growing body of work in this domain.

Going forward, clinicians and researchers should perhaps seek to understand the reasons behind disagreement at a granular level by conducting a robust follow-up analysis with both patients and providers after the initial analysis. Although it is unlikely that this type of analysis could be conducted on a large scale with the general population, a series of targeted studies, focusing on a single condition and within a given subgroups may provide insight into opportunity for improvement that when aggregated can assist in

the creation of more complete reference data and self-report methods that provide more accurate patient self-report.

Finally, given the increasing relevance of measuring agreement between disparate sources of patient clinical data, we believe that more work is needed in developing a standard process and format for reporting underlying data, especially error matrix values. Across all of the articles we reviewed, including articles that were ultimately rejected, there were almost as many ways data values and methods were reported. Understandably, in studies that made use of legacy data sets, identifying and documenting codes and methods used in secondary data sources is challenging, if not impossible. However, in studies explicitly designed to measure inter-rater reliability, greater detail about data lookback periods, completeness of data sources and a list of codes used to identify diagnoses in a reference standard as well as increased transparency about patient self-report methods including survey questions will facilitate measurement and comparison of agreement results. Another theme we found was the inconsistent way questions were often posed to patients. Some researchers asked patients to consider their current conditions, some researchers asked patients about diagnoses within the previous year while others placed no time-limits for patient self-reporting. Additionally, patients were often asked to respond about having vague or ambiguous conditions such as kidney- or heart-related conditions. Greater specificity and more precise language would improve consistency across studies and possible improve agreement across data sources.

Limitations of the analysis

We are limited to the information provided in the articles to determine diagnosis, self-report type and reference standard type. Often, there is a paucity of detail provided about these factors. For example, the term medical record is used frequently to identify reference standard data but no further descriptive information is given. It is not clear whether the term implies electronically captured data or paper-based records which may actually be quite different in nature. Similarly, there is usually little information provided when patients have access to an interviewer as to how much assistance the interviewer actually gave a patient. Therefore, some face-to-face interviews may, in truth, resemble self-administered surveys, whereas, a self-administered survey conducted where patients were able to discuss a question with a medical professional may have actually been more like a face-to-face interview. As such, many of the

variable groupings created for this analysis are more generalized than we would like. Additionally, there may be variations amongst the populations that are not explicitly reported. For example, while respondents in a clinic may not all have been diagnosed a particular condition, we could most likely assume that in some way, many of the respondents have some type of illness present. However, since there is no way for us to distinguish well-visits from diagnostic visits or chronic care, our ability to accurately create like-patient cohorts is also limited. Finally, although a total of 50 unique conditions measured across all cases, many diagnoses only included one or two cases; therefore, the focus of the discussion was limited to diabetes, hypertension, high cholesterol, stroke and myocardial infarction.

REFERENCES

[1] Evans RS. Electronic Health records: then, now, and in the future. Yearb Med Inform. 2016;Suppl 1:S48–61.

[2] Institute of Medicine (US) Roundtable on Value & Science-Driven Health Care Clinical Data as the Basic Staple of Health Learning: Creating and Protecting a Public Good: Workshop Summary. US National Academies Press: Washington, DC, 2010

[3] Horton, D., Bhullar, H., Carty, L., Cunningham, F., Ogdie, A., Sultana, J. and Trifirò, G. (2020). Electronic Health Record Databases. In Pharmacoepidemiology (eds B.L. Strom, S.E. Kimmel and S. Hennessy). doi:10.1002/9781119413431.ch13)

[4] Weiskopf NG, Hripcsak G, Swaminathan S, Weng C. Defining and measuring completeness of electronic health records for secondary use. J. Biomed. Inf. 2013;46:830–836. doi: 10.1016/j.jbi.2013.06.010.

[5] Council for Affordable Quality Healthcare. Defining the Provider Data Dilemma: Challenges, Opportunities, and Call for Industry Collaboration. Washington, DC: CAQH/Manatt Health; 2016. [Accessed January 29, 2020].

[6] Reisman M. EHRs: the challenge of making electronic data usable and interoperable. P T 2017; 42: 572–575.

[7] Hamood, R., Hamood, H., Merhasin, I. et al. A feasibility study to assess the validity of administrative data sources and self-reported information of breast cancer survivors. Isr J Health Policy Res 5, 50 (2016). https://doi.org/10.1186/s13584-016-0111-6

[8] . Dullabh P, Sondheimer N, Katsh E, Evans MA. How patients can improve the accuracy of their medical records. eGEMs, 2 (3) (2014), p. 1080

[9] Rosenman R, Tennekoon V, Hill LG. Measuring bias in self-reported data. Int J Behav Healthc Res. 2011;2:320–32. doi: 10.1504/IJBHR.2011.043414

[10] Yasaitis L.C., Berkman L.F., Chandra A. Comparison of self-reported and Medicare claims-identified acute myocardial infarction. Circulation. 2015;131:1477–1485. doi: 10.1161/CIRCULATIONAHA.114.013829.

[11] Koller K.R., Wilson A.S., Asay E.D., Metzger J.S., Neal D.E. Agreement between self-report and medical record prevalence of 16 chronic conditions in the Alaska earth study. J. Prim. Care Community Health. 2014;5:160–165. doi: 10.1177/2150131913517902.

[12] Muggah E., Graves E., Bennett C., Manuel D.G. Ascertainment of chronic diseases using population health data: A comparison of health administrative data and patient self-report. BMC Public Health. 2013;13:16. doi: 10.1186/1471-2458-13-16.

[13] Smith B, Chu LK, Smith TC, et al. Challenges of self-reported medical conditions and electronic medical records among members of a large military cohort. BMC Med Res Methodol. 2008;8(1):37.

[14] Coughlin SS. Recall bias in epidemiologic studies. J Clin Epidemiol. 1990;43(1):87–91. doi:10.1016/0895-4356(90)90060-3

[15] Lao C.-K., Chan Y.-M., Tong H. H.-Y., Chan A. Underdiagnosis of depression in an economically deprived population in Macao, China. Asia-Pacific Psychiatry. 2016;8(1):70–79. doi: 10.1111/appy.12208.

[16] Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33(1):159–174.

[17] McHugh ML. Interrater reliability: The kappa statistic. Biochem Med (Zagreb) 2012;22:276–282. doi: 10.11613/BM.2012.031

[18] IBM Corp. Released 2019. IBM SPSS Statistics for Windows, Version 26.0. Armonk, NY: IBM Corp.

[19] Eze-Nliam C, Cain K, Bond K, Forlenza K, Jankowski R, Magyar-Russell G, Yenokyan G, Ziegelstein RC (2012) Discrepancies between the medical record and the reports of patients with acute coronary syndrome regarding important aspects of the medical history. BMC Health Serv Res 12(1):78

[20] Kim YY, Park JH, Kang HJ, et al. Level of agreement and factors associated with discrepancies between Nationwide medical history questionnaires and hospital claims data. J Prev Med Public Health. 2017;50(5):294–302.

[21] Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas. 1960;20:37–46.

[22] Conger AJ. Kappa and Rater Accuracy: Paradigms and Parameters. Educ Psychol Meas. 2017;77(6):1019–1047. doi:10.1177/0013164416663277

[23] Johnson EK, Nelson CP. Values and pitfalls of the use of administrative databases for outcomes assessment. J Urol. 2013;190(1):17-18.

[24] Holmes C, Brown M, Hilaire DS, Wright A. Healthcare provider attitudes towards the problem list in an electronic health record: a mixed-methods qualitative study. BMC Med Inform Decis Mak. 2012;12:127. doi:10.1186/1472-6947-12-127

[25] Diaz-Garelli J.-F, Bernstam E. V, MSE, Rahbar M. H, Johnson T. Rediscovering drug side effects: the impact of analytical assumptions on the detection of associations in EHR data. AMIA Summits on Translational Science Proceedings. 2015;2015:51–55.

[26] Wu CS, Lai MS, Gau SS, Wang SC, Tsai HJ. Concordance between patient self-reports and claims data on clinical diagnoses, medication use, and health system utilization in Taiwan. PLoS ONE. 2014;9(12):e112257. doi: 10.1371/journal.pone.0112257.

[27] Peterson KL, Jacobs JP, Allender S, Alston LV, Nichols M. Characterising the extent of misreporting of high blood pressure, high cholesterol, and diabetes using the Australian health survey. BMC Public Health. 2016;16:695.

[28] Tenkorang EY, Sedziafa P, Sano Y, Kuuire V, Banchani E. Validity of self-report data in hypertension research: findings from the Study on Global Ageing and Adult Health. J Clin Hypertens (Greenwich) 2015;17(12):977–984.

[29] Malik AS, Giamouzis G, Georgiopoulou VV, Fike LV, Kalogeropoulos AP, Norton CR, Sorescu D, Azim S, Laskar SR, Smith AL, Dunbar SB, Butler J. Patient perception versus medical record entry of health-related conditions among patients with heart failure. Am J Cardiol. 2011;107(4):569–572. doi: 10.1016/j.amjcard.2010.10.017.

[30] Roberts RO, Bergstralh EJ, Schmidt L, Jacobsen SJ. Comparison of Self-Reported and Medical Record Health Care Utilization Measures. Journal of Clinical Epidemiology. 1996;49:989–995

[31] Fortin M, Haggerty J, Sanche S, Almirall J. Self-reported versus health administrative data: implications for assessing chronic illness burden in populations. A cross-sectional study. CMAJ Open. 2017;5(3):E729-E33.

Appendix A: Descriptives

Diagnosis	N	Mean	Std. Deviation	Range
Anemia	1	0.19	0	-
Angina	3	0.33	0.40	0.78
Anxiety	1	0.44	0	-
Arrhythmia	3	0.36	0.01	0.03
Arthritis	5	0.28	0.09	0.22
Asthma	5	0.50	0.17	0.42
Breast cancer	1	0.81	0	-
Bronchitis/pneumonia	1	0.31	0	-
CAD	1	0.31	0	-
Cancer	5	0.67	0.12	0.33
Cataracts	1	0.21	0	-
Cerebrovascular disease	2	0.53	0.27	0.38
Chronic hepatitis	1	0.36	0	-
Chronic kidney disease	1	0.52	0	-
Chronic low back pain	1	0.20	0	-
Chronic lung disease	1	0.54	0	-
Chronic pulmonary	1	0.18	0	-
Colon cancer	2	0.48	0.18	0.25
Congestive heart failure	6	0.38	0.12	0.34
COPD	5	0.44	0.17	0.36
Coronary heart disease	1	0.77	0	-
Depression	3	0.36	0.14	0.26
Diabetes	27	0.80	0.12	0.61
Emphysema	1	0.82	0	-
Gout	1	0.47	0	-
Heart disease	6	0.50	0.08	0.22
Heart failure	1	0.45	0	-
HepC	1	0.66	0	-
Hiv	1	0.88	0	-
HIV	2	1.00	-	-
Hyperlipidemia	11	0.34	0.24	0.83
Hypertension	28	0.53	0.21	0.75
Impaired cerebral blood flow	1	0.17	0	-
Irritable bowel syndrome	2	0.52	0.04	0.06
Kidney disease	2	0.29	0.22	0.30
Large bowel cancer (CRC)	1	0.75	0	-
Liver disease	3	0.46	0.09	0.16
Lung cancer	1	0.43	0	-
Lung disease	3	0.42	0.19	0.35
Malignancy	1	0.58	0	-
Migraine	1	0.54	0	-
Myocardial infarction	10	0.56	0.18	0.53
Obesity	2	0.41	0.40	0.57
Osteoarthritis	1	0.32	0	-
Osteoporosis	4	0.37	0.13	0.29
Other CVD	1	0.37	0	-
Other heart diseases	1	0.50	0	-
Peptic ulcer disease	2	0.14	0.12	0.16
Peripheral arterial disease	2	0.21	0.15	0.21
Prostate cancer	1	0.78	0	-
Psychiatric disorders	1	0.22	0	-
Pulmonary tuberculosis	1	0.09	0	-
PVD	1	0.32	0	-
Renal/kidney disorders	2	0.41	0.13	0.18
Rheumatoid arthritis	3	0.34	0.30	0.60
Stroke	11	0.50	0.15	0.40
Systemic lupus erythematosus (SLE)	1	0.09	0	-
Thyroid disorder	3	0.71	0.11	0.21
Transient ischemic attack	1	0.10	0	-
Total	**193**	**0.51**	**0.24**	**1.12**

Appendix B: SPSS Analysis: Diagnosis Group Descriptives

Descriptive Statistics

Dependent Variable: kappa

Diagnosis Group	Mean	Std. Deviation	N
Diabetes	.7980233326	.1192468834	27
Hyperlipidemia	.3409386430	.2399145187	11
Hypertension	.5296823986	.2133349631	28
Myocardial infarction	.5593220420	.1847513867	10
Stroke	.4509305512	.1894222084	16
Total	.5753931680	.2406805277	92

Tests of Between-Subjects Effects

Dependent Variable: kappa

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power[b]
Corrected Model	2.252[a]	4	.563	16.220	.000	.427	64.881	1.000
Intercept	22.003	1	22.003	633.958	.000	.879	633.958	1.000
DiagnosisGroup	2.252	4	.563	16.220	.000	.427	64.881	1.000
Error	3.020	87	.035					
Total	35.730	92						
Corrected Total	5.271	91						

a. R Squared = .427 (Adjusted R Squared = .401)

b. Computed using alpha = .05

Parameter Estimates

Dependent Variable: kappa

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval		Partial Eta Squared	Noncent. Parameter	Observed Power[b]
					Lower Bound	Upper Bound			
Intercept	.451	.047	9.682	.000	.358	.544	.519	9.682	1.000
[DiagnosisGroup=Diabetes]	.347	.059	5.905	.000	.230	.464	.286	5.905	1.000
[DiagnosisGroup=Hyperlipidemia]	-.110	.073	-1.507	.135	-.255	.035	.025	1.507	.320
[DiagnosisGroup=Hypertension]	.079	.058	1.349	.181	-.037	.195	.020	1.349	.266
[DiagnosisGroup=Myocardial infarction]	.108	.075	1.443	.153	-.041	.258	.023	1.443	.298
[DiagnosisGroup=Stroke]	0[a]	.	.	.	.	.	.	.	.

a. This parameter is set to zero because it is redundant.

b. Computed using alpha = .05

Pairwise Comparisons

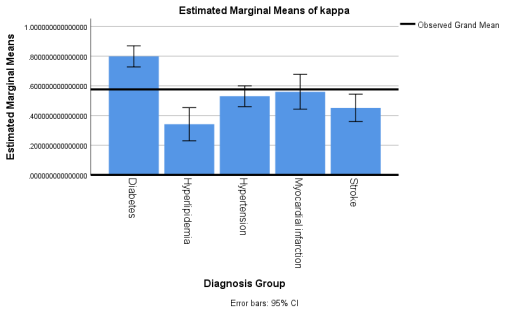
Dependent Variable: kappa

(I) Diagnosis Group	(J) Diagnosis Group	Mean Difference (I-J)	Std. Error	Sig.[b]	95% Confidence Interval for Difference[b]	
					Lower Bound	Upper Bound
Diabetes	Hyperlipidemia	.457*	.067	.000	.266	.648
	Hypertension	.268*	.050	.000	.124	.413
	Myocardial infarction	.239*	.069	.008	.041	.437
	Stroke	.347*	.059	.000	.178	.516
Hyperlipidemia	Diabetes	-.457*	.067	.000	-.648	-.266
	Hypertension	-.189	.066	.054	-.379	.002
	Myocardial infarction	-.218	.081	.084	-.452	.015
	Stroke	-.110	.073	.766	-.320	.100
Hypertension	Diabetes	-.268*	.050	.000	-.413	-.124
	Hyperlipidemia	.189	.066	.054	-.002	.379
	Myocardial infarction	-.030	.069	1.000	-.227	.167
	Stroke	.079	.058	.864	-.089	.246
Myocardial infarction	Diabetes	-.239*	.069	.008	-.437	-.041
	Hyperlipidemia	.218	.081	.084	-.015	.452
	Hypertension	.030	.069	1.000	-.167	.227
	Stroke	.108	.075	.809	-.107	.324
Stroke	Diabetes	-.347*	.059	.000	-.516	-.178
	Hyperlipidemia	.110	.073	.766	-.100	.320
	Hypertension	-.079	.058	.864	-.246	.089
	Myocardial infarction	-.108	.075	.809	-.324	.107

Based on estimated marginal means

*. The mean difference is significant at the .05 level.

b. Adjustment for multiple comparisons: Sidak.



Appendix C: SPSS Analysis: Reference Standard Descriptives

Descriptive Statistics

Dependent Variable: k (calculated)

Ref Std Group	Mean	Std. Deviation	N
BioMx	.3975618952	.2845598622	16
HealthIns Claims	.4577897839	.2097227196	61
Medical Records	.5396257537	.2378023020	111
Registry	.7143072896	.0916817418	5
Total	.5065086127	.2369520235	193

Tests of Between-Subjects Effects

Dependent Variable: k (calculated)

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power[b]
Corrected Model	.672[a]	3	.224	4.191	.007	.062	12.572	.850
Intercept	15.453	1	15.453	288.957	.000	.605	288.957	1.000
RefStdGroup	.672	3	.224	4.191	.007	.062	12.572	.850
Error	10.108	189	.053					
Total	60.294	193						
Corrected Total	10.780	192						

a. R Squared = .062 (Adjusted R Squared = .047)

b. Computed using alpha = .05

Parameter Estimates

Dependent Variable: k (calculated)

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval		Partial Eta Squared	Noncent. Parameter	Observed Power[b]
					Lower Bound	Upper Bound			
Intercept	.714	.103	6.907	.000	.510	.918	.202	6.907	1.000
[RefStdGroup=BioMx]	-.317	.118	-2.673	.008	-.550	-.083	.036	2.673	.758
[RefStdGroup=HealthIns Claims]	-.257	.108	-2.385	.018	-.469	-.044	.029	2.385	.660
[RefStdGroup=Medical Records]	-.175	.106	-1.652	.100	-.383	.034	.014	1.652	.376
[RefStdGroup=Registry]	0[a]	.	.	.	.	.	.	.	.

a. This parameter is set to zero because it is redundant.

b. Computed using alpha = .05

Pairwise Comparisons

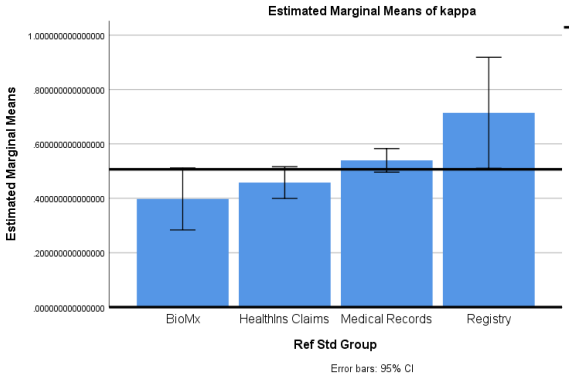
Dependent Variable: kappa

(I) Ref Std Group	(J) Ref Std Group	Mean Difference (I-J)	Std. Error	Sig.[b]	95% Confidence Interval for Difference[b]	
					Lower Bound	Upper Bound
BioMx	HealthIns Claims	-.060	.065	.928	-.233	.112
	Medical Records	-.142	.062	.129	-.306	.022
	Registry	-.317*	.118	.048	-.632	-.002
HealthIns Claims	BioMx	.060	.065	.928	-.112	.233
	Medical Records	-.082	.037	.155	-.180	.016
	Registry	-.257	.108	.104	-.543	.030
Medical Records	BioMx	.142	.062	.129	-.022	.306
	HealthIns Claims	.082	.037	.155	-.016	.180
	Registry	-.175	.106	.469	-.456	.106
Registry	BioMx	.317*	.118	.048	.002	.632
	HealthIns Claims	.257	.108	.104	-.030	.543
	Medical Records	.175	.106	.469	-.106	.456

Based on estimated marginal means

*. The mean difference is significant at the .05 level.

b. Adjustment for multiple comparisons: Sidak.



Appendix D: SPSS Analysis: Self-Report Type Descriptives

Descriptive Statistics

Dependent Variable: kappa

Self Rpt Group	Mean	Std. Deviation	N
F2F/Phone	.4901866837	.2313579286	88
SA/Mail	.5201879436	.2417919037	105
Total	.5065086127	.2369520235	193

Tests of Between-Subjects Effects

Dependent Variable: kappa

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power[b]
Corrected Model	.043[a]	1	.043	.767	.382	.004	.767	.140
Intercept	48.874	1	48.874	869.421	.000	.820	869.421	1.000
SelfRptGroup	.043	1	.043	.767	.382	.004	.767	.140
Error	10.737	191	.056					
Total	60.294	193						
Corrected Total	10.780	192						

a. R Squared = .004 (Adjusted R Squared = -.001)

b. Computed using alpha = .05

Parameter Estimates

Dependent Variable: kappa

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval		Partial Eta Squared	Noncent. Parameter	Observed Power[b]
					Lower Bound	Upper Bound			
Intercept	.520	.023	22.482	.000	.475	.566	.726	22.482	1.000
[SelfRptGroup=F2F/Phone]	-.030	.034	-.876	.382	-.098	.038	.004	.876	.140
[SelfRptGroup=SA/Mail]	0[a]	.	.	.	.	.	.	.	.

a. This parameter is set to zero because it is redundant.

b. Computed using alpha = .05

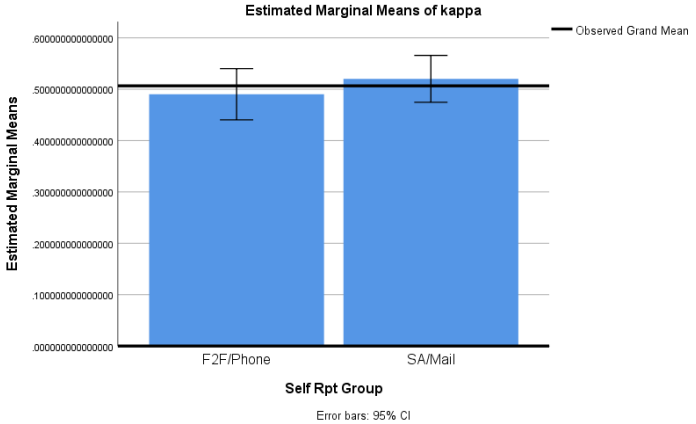
Pairwise Comparisons

Dependent Variable: kappa

(I) Self Rpt Group	(J) Self Rpt Group	Mean Difference (I-J)	Std. Error	Sig.[a]	95% Confidence Interval for Difference[a]	
					Lower Bound	Upper Bound
F2F/Phone	SA/Mail	-.030	.034	.382	-.098	.038
SA/Mail	F2F/Phone	.030	.034	.382	-.038	.098

Based on estimated marginal means

a. Adjustment for multiple comparisons: Sidak.



Appendix E: SPSS Analysis: Region Descriptives

Descriptive Statistics

Dependent Variable: kappa

Region	Mean	Std. Deviation	N
AsiaPac	.4471795196	.2220621975	40
Europe	.5391819140	.2293169881	29
North America	.5283606922	.2380225189	114
Other	.3999587049	.2672510402	10
Total	.5065086127	.2369520235	193

Tests of Between-Subjects Effects

Dependent Variable: kappa

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power[b]
Corrected Model	.340[a]	3	.113	2.050	.108	.032	6.150	.520
Intercept	21.788	1	21.788	394.432	.000	.676	394.432	1.000
Region	.340	3	.113	2.050	.108	.032	6.150	.520
Error	10.440	189	.055					
Total	60.294	193						
Corrected Total	10.780	192						

a. R Squared = .032 (Adjusted R Squared = .016)

b. Computed using alpha = .05

Parameter Estimates

Dependent Variable: kappa

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval		Partial Eta Squared	Noncent. Parameter	Observed Power[b]
					Lower Bound	Upper Bound			
Intercept	.400	.074	5.381	.000	.253	.547	.133	5.381	1.000
[Region=AsiaPac]	.047	.083	.568	.571	-.117	.211	.002	.568	.087
[Region=Europe]	.139	.086	1.615	.108	-.031	.309	.014	1.615	.362
[Region=North America]	.128	.078	1.656	.099	-.025	.281	.014	1.656	.378
[Region=Other]	0[a]	.	.	.	.	.	.	.	.

a. This parameter is set to zero because it is redundant.

b. Computed using alpha = .05

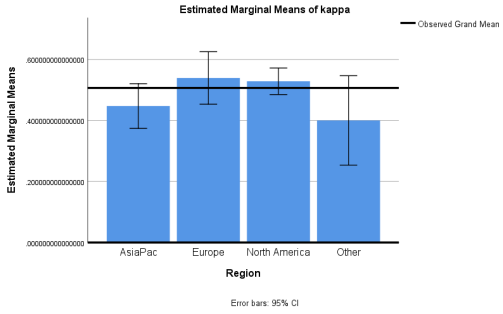
Pairwise Comparisons

Dependent Variable: kappa

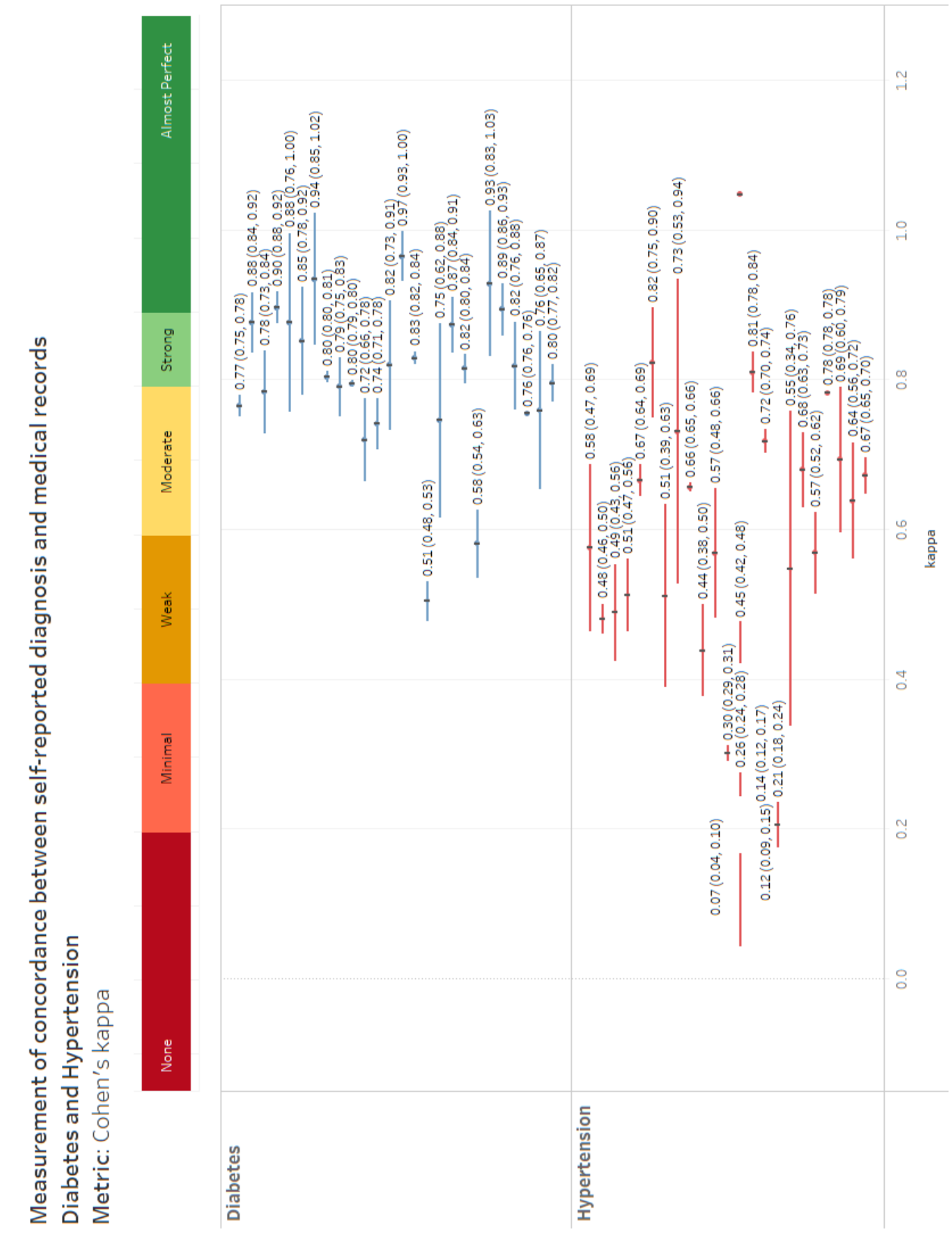
(I) Region	(J) Region	Mean Difference (I-J)	Std. Error	Sig.[a]	95% Confidence Interval for Difference[a]	
					Lower Bound	Upper Bound
AsiaPac	Europe	-.092	.057	.661	-.245	.061
	North America	-.081	.043	.370	-.196	.034
	Other	.047	.083	1.000	-.174	.269
Europe	AsiaPac	.092	.057	.661	-.061	.245
	North America	.011	.049	1.000	-.120	.141
	Other	.139	.086	.647	-.091	.369
North America	AsiaPac	.081	.043	.370	-.034	.196
	Europe	-.011	.049	1.000	-.141	.120
	Other	.128	.078	.596	-.078	.335
Other	AsiaPac	-.047	.083	1.000	-.269	.174
	Europe	-.139	.086	.647	-.369	.091
	North America	-.128	.078	.596	-.335	.078

Based on estimated marginal means

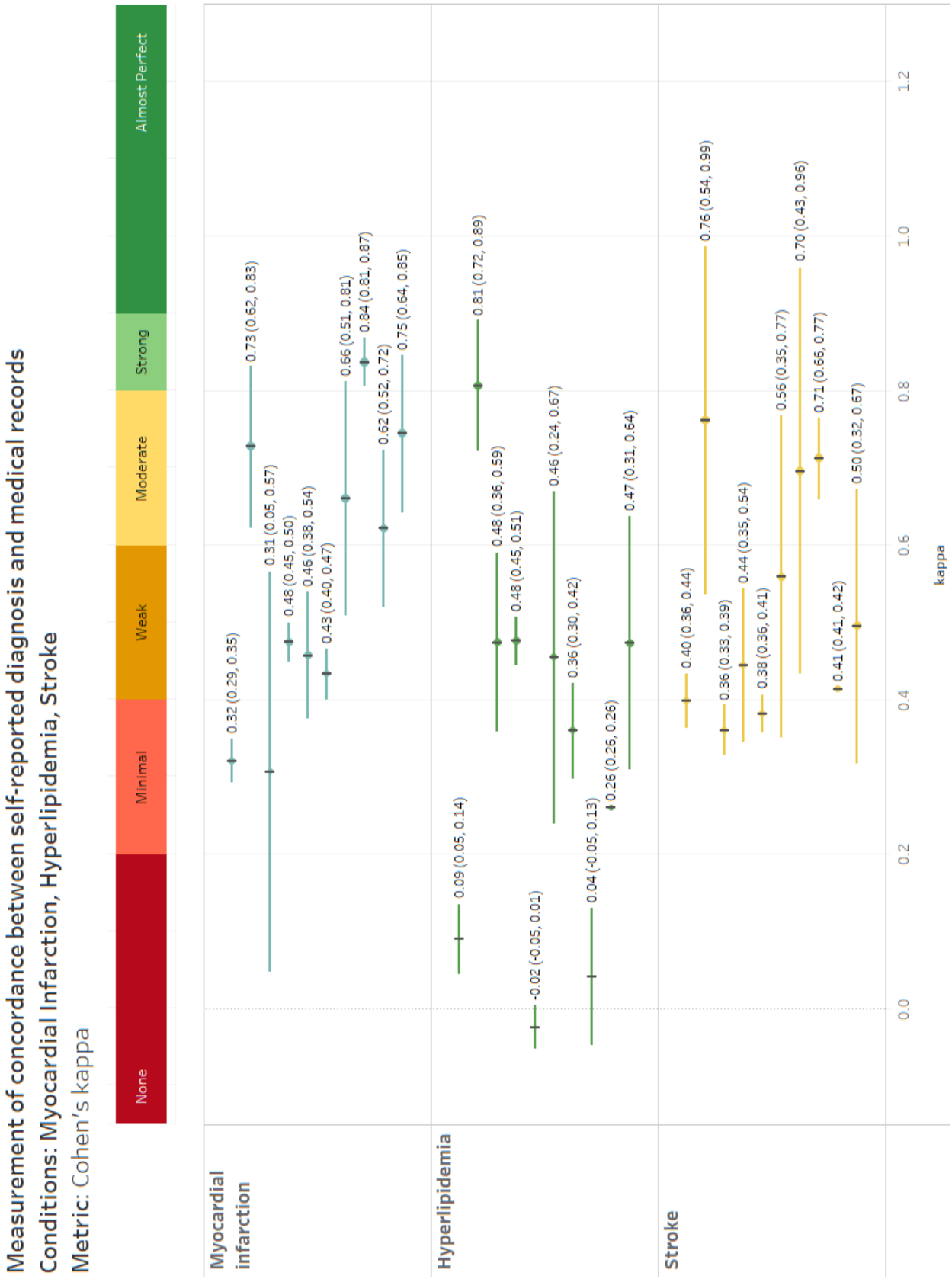
a. Adjustment for multiple comparisons: Bonferroni.



Appendix F: Agreement Summary: Diabetes and Hypertension



Appendix G: Agreement Summary: Myocardial Infarction, Hyperlipidemia, Stroke



Appendix H: Articles Used in Literature Review and Meta-Analysis

Author	Title	PMID
Tormo, M. J., et al.	Validation of self diagnosis of high blood pressure in a sample of the Spanish EPIC cohort: overall agreement and predictive values. EPIC Group of Spain	10746117
Navarro, C., et al.	Validity of self reported diagnoses of cancer in a major Spanish prospective cohort study	16790831
Miller, D. R., et al.	Patients' self-report of diseases in the Medicare Health Outcomes Survey based on comparisons with linked survey and medical data from the Veterans Health Administration	18360178
Walitt, B. T., et al.	Validation of self-report of rheumatoid arthritis and systemic lupus erythematosus: The Women's Health Initiative	18398940
Gravely-Witte, S., et al.	Cardiologists' charting varied by risk factor, and was often discordant with patient report	18411042
Huerta, J. M., et al.	Accuracy of self-reported diabetes, hypertension and hyperlipidemia in the adult Spanish population. DINO study findings	19232187
Barber, J., et al.	Measuring morbidity: self-report or health care records?	20019091
Horton, M., et al.	Validation of a self-report comorbidity questionnaire for multiple sclerosis	20551692
Klein, B. E., et al.	Self- and registry-reported cancer in a population-based longitudinal study	21066931
Malik, A. S., et al.	Patient perception versus medical record entry of health-related conditions among patients with heart failure	21185003
Eze-Nliam, C., et al.	Discrepancies between the medical record and the reports of patients with acute coronary syndrome regarding important aspects of the medical history	22448755
Gure, T. R., et al.	Predictors of self-report of heart failure in a population-based survey of older adults	22592753
Muggah, E., et al.	Ascertainment of chronic diseases using population health data: a comparison of health administrative data and patient self-report	23302258
Leong, A., et al.	Estimating the population prevalence of diagnosed and undiagnosed diabetes	23656982
Teh, R., et al.	Agreement between self-reports and medical records of cardiovascular disease in octogenarians	23860185
Comino, E. J., et al.	Validating self-report of diabetes use by participants in the 45 and Up Study: a record linkage study	24245780
Jackson, J. M., et al.	Validity of diabetes self-reports in the Women's Health Initiative	24496083
Sakshaug, J. W., et al.	Identifying diabetics in Medicare claims and survey data: implications for health services research	24693862
Bai, J. R., et al.	Concordance between medical records and interview data in correctional facilities	24716525
Sridharan, S., et al.	A self-report comorbidity questionnaire for haemodialysis patients	25135668
Lujic, S., et al.	Variation in the recording of common health conditions in routine hospital data: study using linked survey and administrative data in New South Wales, Australia	25186157
Wu, C. S., et al.	Concordance between patient self-reports and claims data on clinical diagnoses, medication use, and health system utilization in Taiwan	25464005
Jackson, C. A., et al.	Moderate agreement between self-reported stroke and hospital-recorded stroke in two cohorts of Australian women: a validation study	25613556
Yasaitis, L. C., et al.	Comparison of self-reported and Medicare claims-identified acute myocardial infarction	25747935
Yuan, X., et al.	Validity of self-reported diabetes among middle-aged and older Chinese adults: the China Health and Retirement Longitudinal Study	25872937
Tenkorang, E. Y., et al.	Validity of Self-Report Data in Hypertension Research: Findings From The Study on Global Ageing and Adult Health	26224341
Jiang, L., et al.	Concordance between Self-Reports and Medicare Claims among Participants in a National Study of Chronic Disease Self-Management Program	26501047
Lovaas, K. F., et al.	Feasibility of using self-reported patient data in a national diabetes register	26666413
Vigen, C., et al.	Validation of self-reported comorbidity status of breast cancer patients with medical records: the California Breast Cancer Survivorship Consortium (CBCSC)	26797455
Chun, H., et al.	Accuracy of Self-reported Hypertension, Diabetes, and Hypercholesterolemia: Analysis of a Representative Sample of Korean Older Adults	27169009
Peterson, K. L., et al.	Characterising the extent of misreporting of high blood pressure, high cholesterol, and diabetes using the Australian Health Survey	27484257
Eliassen, B. M., et al.	Validity of self-reported myocardial infarction and stroke in regions with Sami and Norwegian populations: the SAMINOR 1 Survey and the CVDNOR project	27903562
Hamood, R., et al.	A feasibility study to assess the validity of administrative data sources and self-reported information of breast cancer survivors	27980719
Ye, F., et al.	Comparison of Patient Report and Medical Records of Comorbidities: Results From a Population-Based Cohort of Patients With Prostate Cancer	28208186
Borlee, F., et al.	Spirometry, questionnaire and electronic medical record based COPD in a population survey: Comparing prevalence, level of agreement and associations with potential risk factors	28273094
Fortin, M., et al.	Self-reported versus health administrative data: implications for assessing chronic illness burden in populations. A cross-sectional study	28947426
Kim, Y. Y., et al.	Level of Agreement and Factors Associated With Discrepancies Between Nationwide Medical History Questionnaires and Hospital Claims Data	29020761
Wagaw, F., et al.	Linking Data From Health Surveys and Electronic Health Records: A Demonstration Project in Two Chicago Health Center Clinics	29346063
Hoffmann, J., et al.	How do patients with diabetes report their comorbidities? Comparison with administrative data	29750054
Chiu, C. J., et al.	National health data linkage and the agreement between self-reports and medical records for middle-aged and older adults in Taiwan	30509280