210 Haag Hall
5100 Rockhill Road
Kansas City, MO 64110
(816) 235-1314
*cei.umkc.edu*

Center for Economic Information
Working Paper 1702-01
July 21, 2017

# *An Iterative Approach to the Parcel Level Address Geocoding of a Large Health Dataset to a Shifting Household Geography*

by
Ben Wilson and Neal Wilson

## Abstract

This article details an iterative process for the address geocoding of a large collection of health encounters (n = 242,804) gathered over a 13 year period to a parcel geography which varies by year. This procedure supports an investigation of the relationship between basic housing conditions and the corresponding health of occupants. Successful investigation of this relationship necessitated matching individuals, their health outcomes and their home environments. This match process may be useful to researchers in a variety of fields with particular emphasis on predictive modeling and up-stream medicine.

# *An Iterative Application of Centerline and Parcel Geographies for Spatio-temporal Geocoding of Health and Housing Data*

Ben Wilson & Neal Wilson

## 1 Introduction

This article explains an iterative geocoding process for matching address level health encounters (590,058 asthma and well child encounters ) to heterogeneous parcel geographies that maintains the spatio-temporal disaggregation of both datasets (243,260 surveyed parcels). The development of this procedure supports the objectives outlined in the U.S. Department of Housing and Urban Development funded Kansas City – Home Environment Research Taskforce (KC-Heart). The goal of KC-Heart is investigate the relationship between basic housing conditions and the health outcomes of child occupants. Successful investigation of this relationship necessitates matching individuals, their health outcomes and their home environments. While this geocoding process is designed for the specific needs of the KC-Heart data it has the benefit of producing a disaggregated collection of spatio-temporal shapefiles that suit the exploration and development of additional research questions to those explicit in the founding of KC-Heart. This research flexibility is created by generating several distinct geographies (centerline, and parcel level point layers, and larger administrative polygon geographies) as reference data in the geocoding process.  The resultant set of geocoded shapefiles facilitate the examination of a variety of potential confounders regardless of their availability at specific geographic resolutions, while also adhering to traditional standards for geocoding in terms of positional accuracy, completeness, and repeatability.  We conclude that the increased accuracy and flexibility of the data generated through this process not only supports the objectives of KC-Heart, but are recommended as a standard practice where address level encounter data is available.

## 2 GIS Spatial Analysis and Research

### *2.1 Background*

KC-Heart is focused on public health and housing. We argue the method outlined below is useful to a range of research and policy applications. The interdisciplinary application most germane to KC-Heart is “preventative medicine”. Preventative medicine is a holistic approach that is recognized as a cost conscious and effective treatment strategy for chronic diseases like asthma (Matsui 2016). The aim of preventative medicine is to address the environmental risk factors for chronic disease and manage the treatment of existing symptoms of disease. This represents a reasonable approach to medicine in light of what is known about the connection between the lived environment and disease. For example evidence of the importance of where you live is presented in research linking incidents of childhood asthma with proximity to highways and railroads (Brauer et al 2002, Schuch et al 2016), links between lead exposure in daily life and elevated childhood blood lead levels (Rabito et al 2006,)), household conditions and poverty (Rauh et al. 2008), and food deserts and childhood allergies (Humphries et al. 2015).

Doing this sort of interdisciplinary GIS work brings up several issues with regard to data identification. First, the process of identifying which are the relevant data to the research question requires the expertise of doctors, epidemiologists, statisticians, public health workers and economists. Each of these disciplines brings a distinct focus to the problem that, when combined, constitutes a more complete understanding of which data are needed to model the problem than could be developed independently. Second, this more expansive listing of data necessitates collaboration among institutions where the needed datasets are typically siloed. These issues, of data and expertise, converge in the geocoding process. An example from KC-Heart makes these points clear. Exactly which data are relevant to modeling the social determinants of asthma are not clear a priori. The health encounter data, demographic data, housing data, environmental data, and census data relevant to modeling asthma are typically kept at separate institutions that specialize in these particular areas. Finally, these data sets do not originate in the GIS format. Thus, the process of migrating the data such that they become spatially organized is a crucial intermediate step that takes place after collection but before research questions can be examined.

This example highlights the importance of approaching issues of public health by way of a consortium of experts to identify relevant risk factors, pool data sources and structure the inquiry consistent with identified best practices.[1] Although every disease and economic phenomena exhibit spatial characteristics, determining their extent and significance requires care and multiple perspectives. The ability of GIS to clearly structure, associate, and display information regarding a particular health outcome in a common geography makes it a bridge between specialists and a natural framework for

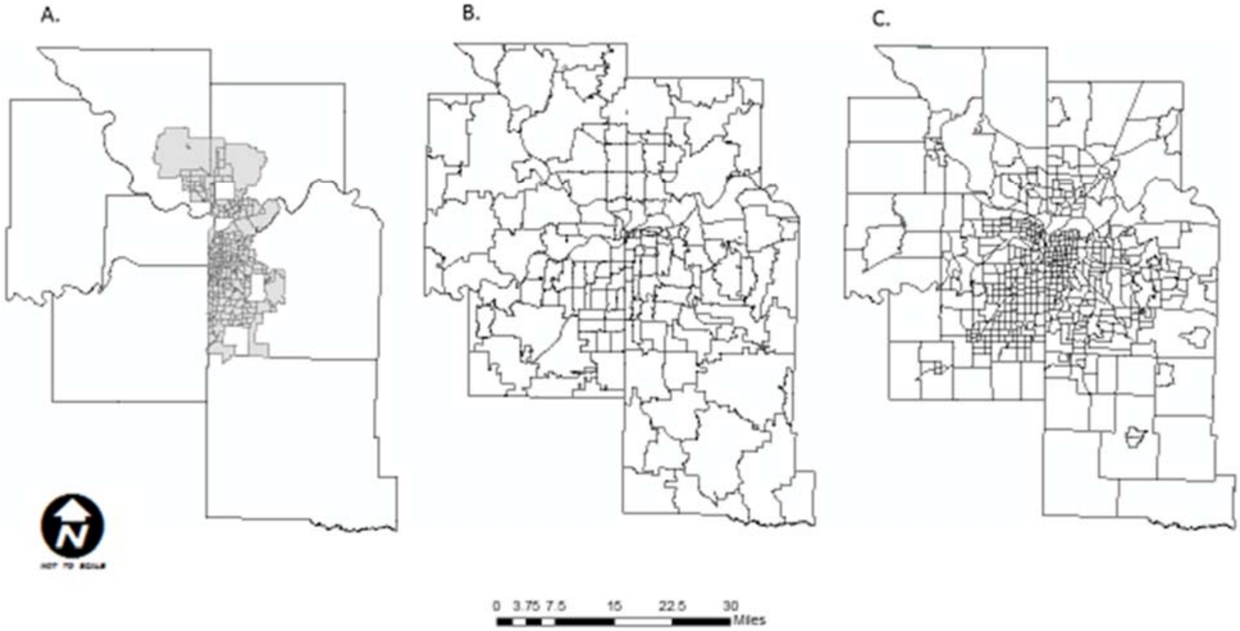
[1] For example, a best practice from economic analysis is to use as much of the available data as possible. Our process is designed with this in mind.

building new knowledge about the relationships between health and the lived environment.

### *2.2 Spatial Issues*

Due to its role as the point of convergence at which diverse data sets meet, every study using GIS relies on the validity of its data geocoding (or address matching) procedure as the basis for its results. Accordingly, a critical literature has developed which discusses the relative merits of geocoding to various levels of aggregation (Openshaw 1984; Zandbergen 2008; Briant 2010; Jacquez 2012). Some research uses an aggregated administrative scale such as commuting zone, zip code or neighborhood geographies to perform their analysis (Chetty et al. 2014; Chetty et al. 2016). The use of large administrative geographies has the advantages of being cost effective with regard to the geocoding process, high match rates and data availability (Rushton 2006). The disadvantages of using such geographies are well documented with regard to the introduction of placement error into epidemiological exposure models as well as the ecological fallacy (Susser 1994; Jacquez 2012; Edwards et al. 2014). Address matching on the basis of a street centerline geography is recognized as being superior to the aggregated administrative boundary, however this geography too introduces an unnecessary degree of placement error (Rushton 2006; Edwards et al. 2014). For locational accuracy it is well recognized that the best level of geography to use for epidemiology and public health studies is the parcel centroid (Rushton 2006; Manson et al 2009; Jacquez 2012). It is noted that accuracy introduces cost concerns, data loss from lower match rates and potential privacy issues for patients (McElroy et. al 2003; Rushton 2006).

Figure1. Distinct administrative boundaries for the Metropolitan Kansas City area



Three maps of the Kansas City Metropolitan Statistical Area (MSA) are presented in Figure 1 to demonstrate this point regarding relevant polygon geographies. These maps show the MSA disaggregated into different administrative polygon geographies, the county, the City of Kansas City Missouri and its corresponding neighborhoods in map 1.A, zip codes in map 1.B, and census tracts in map1.C. It is clear that these three geographies represent the underlying MSA in different and mutually inconsistent ways. Thus any analysis done on the basis of geocoding these administrative boundaries becomes a function of the boundaries themselves and not simply the phenomena under examination. This limits the ability of the analyst to make claims about potential underlying spatial processes. In Figure 2, the resolution is further disaggregated to the neighborhood (map 2.A), the street centerline (map 2.B), and parcel geographies (map 2.C). While, centerline matches are vastly more accurate than neighborhoods, parcel matches are more precise yet. Figure 3 demonstrates two problems endemic to centerline matches that are well discussed in the literature. (Rushton 2006; Jacquez 2012) Map 3.A demonstrates that matching to this geography tends to bunch encounters at the end of centerline segments while map 3.B shows the centerline geography to misalign addresses in the absence of end-of-line-segment clustering. ,

Figure 2. Three Views of the Lykins Neighborhood

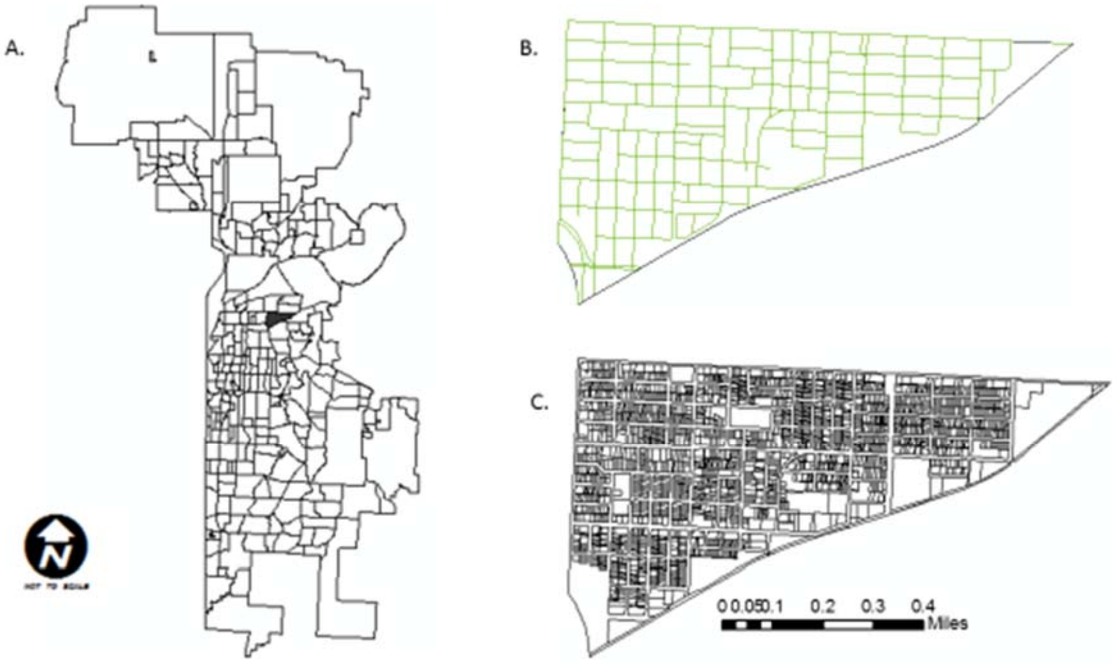
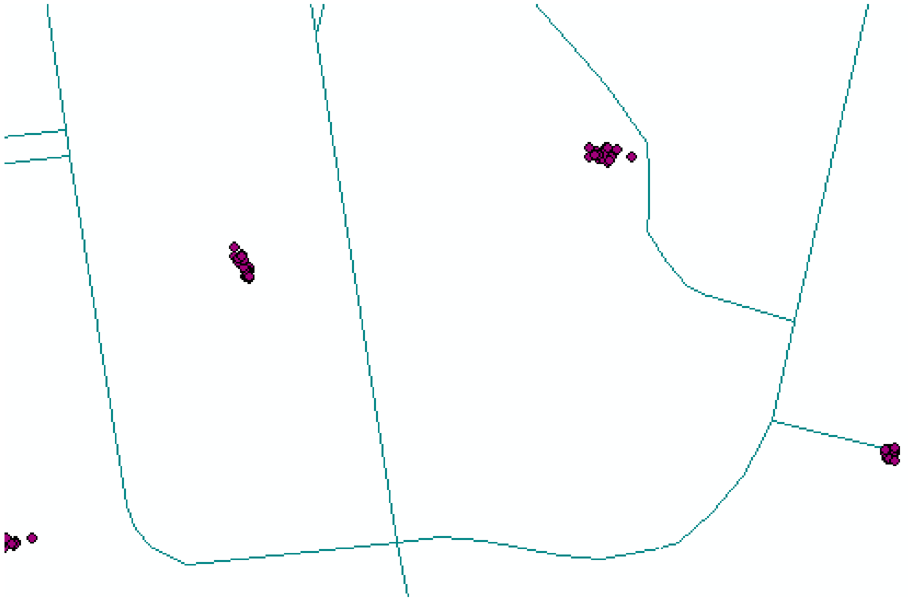


Figure 3.Clustered Matches along Centerline Segments



Despite its superiority for analysis there are tradeoffs to consider when geocoding to a more precise address geography. The most frequently discussed is that match rates tends to fall as the level of specificity becomes more exact (Dearwent et al 2001; Rushton 2006; Zandbergen 2007; Manson et al 2009) Match rates fall at the parcel level for three composite reasons. Matching addresses to the parcel requires greater accuracy in the address file to be matched than is the case in centerline matching. Centerline matches are based on a range of addresses while parcels have a unique and exact address. In practice this causes secondary addresses (i.e. 2002 ½ Main St) associated with a parcel to fail to match. Frequently there are issues with the availability of data at finer resolutions that makes geocoding to the parcel geography irrelevant. Most publicly available data on health, income, and demographics is not available at the household or parcel level. Matching some but not all data to the parcel geography creates methodological complications that can be avoided by choosing a higher level of aggregation for matching. The Health Insurance Portability and Accountability Act (HIPPA) created a set of guidelines for researchers regarding the confidentiality of study subjects that, in some sense, discourages address level analysis. These issues contribute to preventative medicine, predictive analysis, and policy analysis being done at the higher levels of geography despite the recognized superiority of analysis at the parcel geography. These larger levels of aggregation, however, greatly simplify the space under analysis and thus makes it difficult to identify and explore potential spatial processes. Hopefully visual inspection of the maps above brings into view the related ideas of the modifiable aerial unit problem (MAUP) and ecological fallacy. These two related ideas call into question the validity of those analysis performed at an inappropriately general level of spatial resolution.[2]

### *2.3 Temporal Issues*

Just as important as spatial considerations are temporal issues in GIS work. There are two principle classes of temporal issues discussed in the GIS literature (Beale et al 2008; Abellan 2008; Mason et al 2009, Kwan 2012). One class of issues deals with the temporal stability of environmental measurements, the other deals with the mobility of subjects during the time of measure. It has been suggested that these classes are aspects of a singular "Uncertain Geographic Context Problem" (UGCoP) and can be addressed using the same techniques (Kwan 2012). Nevertheless, considering these issues one at a time, it is clear that health relevant observations take place at points in time while the value of what is being measured changes over time. Consider the observed characteristics of a parcel geography at time *t*, this observation does not tell us what the value of that same characteristic will be at time $t + 1$ nor the change over time. There exists a tradeoff between exactitude and the availability of temporal data just as there is

[2] For more on the relationship between MAUP, the ecological fallacy and our project please see WP 1703-01

with regard to spatial analysis. For this reason researchers must be careful and explicit regarding temporal context when combining observations in GIS. Related to this uncertain geographic context problem, is the behavior of the patient under study. There is frequently an assumption made in GIS work that the patient's location, and thus their relevant health context, is more stable than it actually is. Households can change addresses without warning and the regular movement of individual subjects during the day may confound the reliance on such relevant geographies as neighborhoods for contextual influence on human health (Kwan 2012). Giving credence to temporal issues requires an emphasis on modeling health in space-time rather than just space.

### *2.4 Overview*

This article addresses many of the issues discussed above by explaining an iterative process for geocoding a large volume of address data to heterogeneous temporal geographies. This geocoding procedure is reported as an aspect of the U.S. Department of Health and Human Services funded Kansas City – Home Environment Research Taskforce (KC-Heart). The goal of KC-Heart is to advance knowledge about the relationship between housing conditions and the corresponding health of occupants with regard to the health outcomes of childhood asthma, childhood lead poisoning and childhood injury. The success of [phase 1 of] KC-Heart is based on the legitimacy of the address matching procedure that associates the health of individuals to their home environments. This geocoding process was designed to match 590,058 Asthma and Well Child health encounters in the Kansas City metropolitan area to observed housing conditions of 243,260 parcels over 13 years. The geocoding process described below is designed to be extended with regard to KC-Heart to 238,756 observations of pediatric lead poisoning and 1,172,606 incidents of childhood injury. The design of an address match process for these large data sets is complicated in KC-heart due to the year to year variation in the target parcel geography. The iterative geocoding process described below is designed for the specific challenges of the KC-Heart data and, more broadly, to facilitate the examination of environmental health and preventative medicine from a multitude of spatio-temporal aggregations.

The selection of a geocoding procedure needs to be based upon the intended use and research questions associated with the data in question (Rushton 2006, Kwan 2012). Aim 1 of KC-Heart is built on matching individual health records (nearly 600,000 of them) to specific household locations (nearly 245,000 of these). Matching these records to the parcel geography is preferred for reasons discussed above and to take advantage of observations of exterior housing conditions at the parcel level. This match process is complicated by the fact that the observations of parcel level housing condition are comprised by 33 distinct survey files. A match procedure could have been completed by simply developing distinct address locator files for each of the survey parcel geographies. However, this process would have yielded no relevant indication of match rate and consequently resulted in the loss of health encounter data. By applying an iterative process utilizing both centerline and polygon sourced address locators, a meaningful match rate (discussed below) is created and the maximum amount of data is available for the intended analysis. In addition, the iterative approach generates an annualized collection of point shapefiles across both the centerline and polygon geographies to construct a disaggregated spatio-temporal geodatabase of housing and health information. This allows for the investigation of housing conditions on health and the tracking on individual patients over time without regard to their residence in a surveyed parcel.

Geocoding procedures are judged with regard to the identified criteria of completeness, positional accuracy, and repeatability (Zandbergen 2009; Jacquez 2012). To these standard criteria we add a consideration of cost and flexibility of geocoded data. Completeness is the percentage of relevant records that can be reliably geocoded. Sometimes called the hit rate or match rate, calculating this statistic presents some issues that will be discussed in the following sections. Positional accuracy accounts for the relationship between the geocoded location of the address in question and its “true” location. Repeatability indicates how dependent the geocoding results are on the particular software, skills and sensitivity of the analyst. (Whitsel et al. 2004; Edwards et al. 2014) Cost can be a critical consideration as address geocoding can be voracious in its consumption of time and money while flexibility can be thought of as subset of the general usefulness of data. In the following sections, it is demonstrated how an iterative approach to geocoding not only satisfies these criteria, but also supplies a collection of disaggregated shapefiles to facilitate wider research inquiries.

## 3 Method

Several institutions came together to form KC-Heart (Children's Mercy Hospital (CM), Mid-America Regional Council (MARC), the Kansas City, Missouri Health Department, and the Center for Economic Information (CEI) at the University of Missouri Kansas City) with each contributing data to the geocoding process described below. These are two important points regarding the handling of data in KC-heary. The first is regarding the HIPPA required Internal Review Board (IRB) Protocol. Considerable time and effort was saved through Children's Mercy Hospital's origination of the KC-Heart Program and the use of their IRB as the anchor to which the associated institutions request to rely, this expedited the research process and trimmed potentially redundant labor. Second, it is recognized that the data which goes into this sort of analysis is dispersed. There are significant institutional and legal issues that make the assembly of relevant data difficult. No analysis would have taken place without the data sharing agreements between institutions and the data consortium housed at Children's Mercy Hospital. The health data, project geographies and housing conditions data provided by these partners in KC-Heart are described briefly before the discussion of the geocoding process.[3]

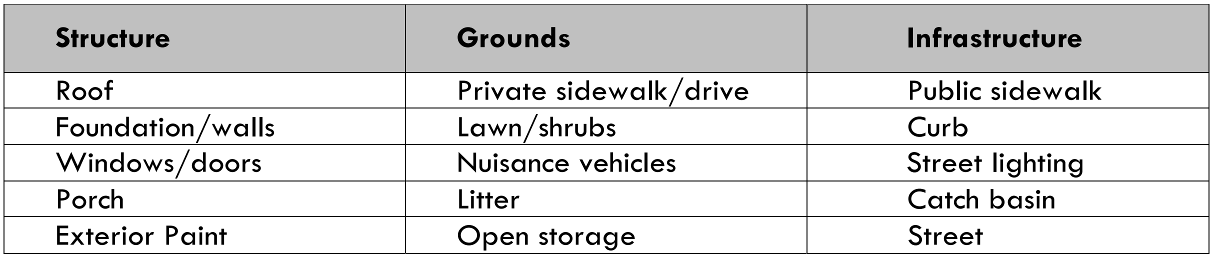
The individual KC-Heart data is a collection of health encounter and claims data selected on the basis of ICD-9 diagnosis codes for visits at Children's Mercy Hospital between 2000 and 2015. Each patient in the Children's Mercy system receives a unique medical record number (MRN) while each visit to the hospital is given a unique account number (AcctNum). These two identifiers allow researchers to analyze patient demographics and differentiate between incidences of the severity of multiple encounters. It is thus possible to track the frequency of visits, differentiate between visits, and follow the same individual through multiple visits over the complete time frame of the investigation. Other Attribute data, such as patient's demographic and treatment information were joined by the unique identifiers after the matching process was completed.

The project geography used in the geocoding process included both line and polygon shapefiles. The critical line shapefile is the greater Kansas City Metropolitan area street centerline file. This covers an area extending into seven counties and crossing into two states. This centerline geography is maintained by MARC and used to coordinate regional development initiatives. Throughout the project, centerline matches were based on MARC's 2014 street centerline file. The parcel geography utilized is based on 33 neighborhood housing condition surveys (NHCS) carried out by the Center for Economic Information (CEI) between 2000 and 2012[4]. Each year of the survey examined the housing conditions in a different subset of the metropolitan parcel geography. The CEI parcel geography contains 15 different components measured using a five point scale developed using Kansas City, MO housing codes as a consistent

[3] For more information about the assembly and association of environmental factors with regard to the intensity of asthma in our study please see WP 1704-01 .

[4] For more about the NHCS please see CEI working paper 1701-01.

benchmark. Complementing the housing conditions data are several categorical variables that further describe the parcel, such as structure type and land use codes. The five point scale for each of the 15 characteristics with corresponding value criteria is as follows: 5 - Excellent, 4 - Good, 3 - Sub-standard, 2- Seriously deteriorated, 1 - Severely deteriorated. The 15 characteristics are broken into three general categories – structure characteristics, grounds characteristics, and city infrastructure characteristics. Figure 5 shows the components of each of the general groups.



| Structure | Grounds | Infrastructure |
|---|---|---|
| Roof | Private sidewalk/drive | Public sidewalk |
| Foundation/walls | Lawn/shrubs | Curb |
| Windows/doors | Nuisance vehicles | Street lighting |
| Porch | Litter | Catch basin |
| Exterior Paint | Open storage | Street |

Figure 4: Observational components of NHCS .

Also included in the parcel file are multiple types of address identifier such as KIVA-pin, County-ID and a text filed with a legal description of the parcel. The “Legal” field is particularly important as it lists the secondary addresses that correspond to the parcel. This data is useful with regard to apartment buildings where there may be multiple (sometimes upwards of 20) addresses associated with a single parcel. Additional polygon shapefiles of higher orders such as zip codes, census block, and neighborhood levels of geography are provided by the CEI. Each of these shapefiles contributes to the geocoding process and the construction of a geodatabase to be shared among the research group.

The iterative address match process embodies several distinct steps. This gives a general sketch of the process while a more thorough accounting follows. First a centerline geocoding of the health data is completed using the default settings on the address locator program in ArcMAP. Then a second centerline geocoding of just the unmatched encounters to a moderately lower setting takes place. These first two levels of geocoding are used to identify those encounters that are in proximity to the NHCS and provide a spatial distribution of the entire health encounters sample. The matched encounters from steps one and two are subset by proximity to the NHCS and re-geocoded, this time to the parcel geography based on the housing conditions survey. Finally a manual geocoding to the parcel level is undertaken of only the unmatched encounters from step three. It is only those encounters which were geocoded to the parcel either through the automated matching process or the manual rematch process that were passed on for further analysis vis-a-vis housing conditions consistent with the design of the study.

The detailed geocoding process was designed in accordance with the criteria of completeness, positional accuracy and repeatability in accordance with the best practices for geocoding for medical research (McElroy et al 2003; Rushton et al 2006; Manson et al 2008; Jacquez 2012; Sonderman et al 2012). A detailed accounting of the geocoding process is somewhat more involved than the general sketch above. Due to changes in geography between each housing condition survey and the large volume of data to be matched we separated the CMH supplied health encounters by year and by state for the initial centerline matches, then by year and by NHCS program the final parcel matches. To make sure we are using contemporaneous observations we match health encounters only to the NHCS parcels surveyed during the same year as the health encounter.

Keeping with the principle of reproducibility we follow a consistent process such that we are able to merge data from different surveys and different years with confidence that they are generated using a consistent method. Below we describe the process for a single year of a single housing condition survey. Figure 5 is a schematic representation of the process. Appendix 1 contains the exact process for the year 2012 which is representative of the process followed in each of the years of our analysis.

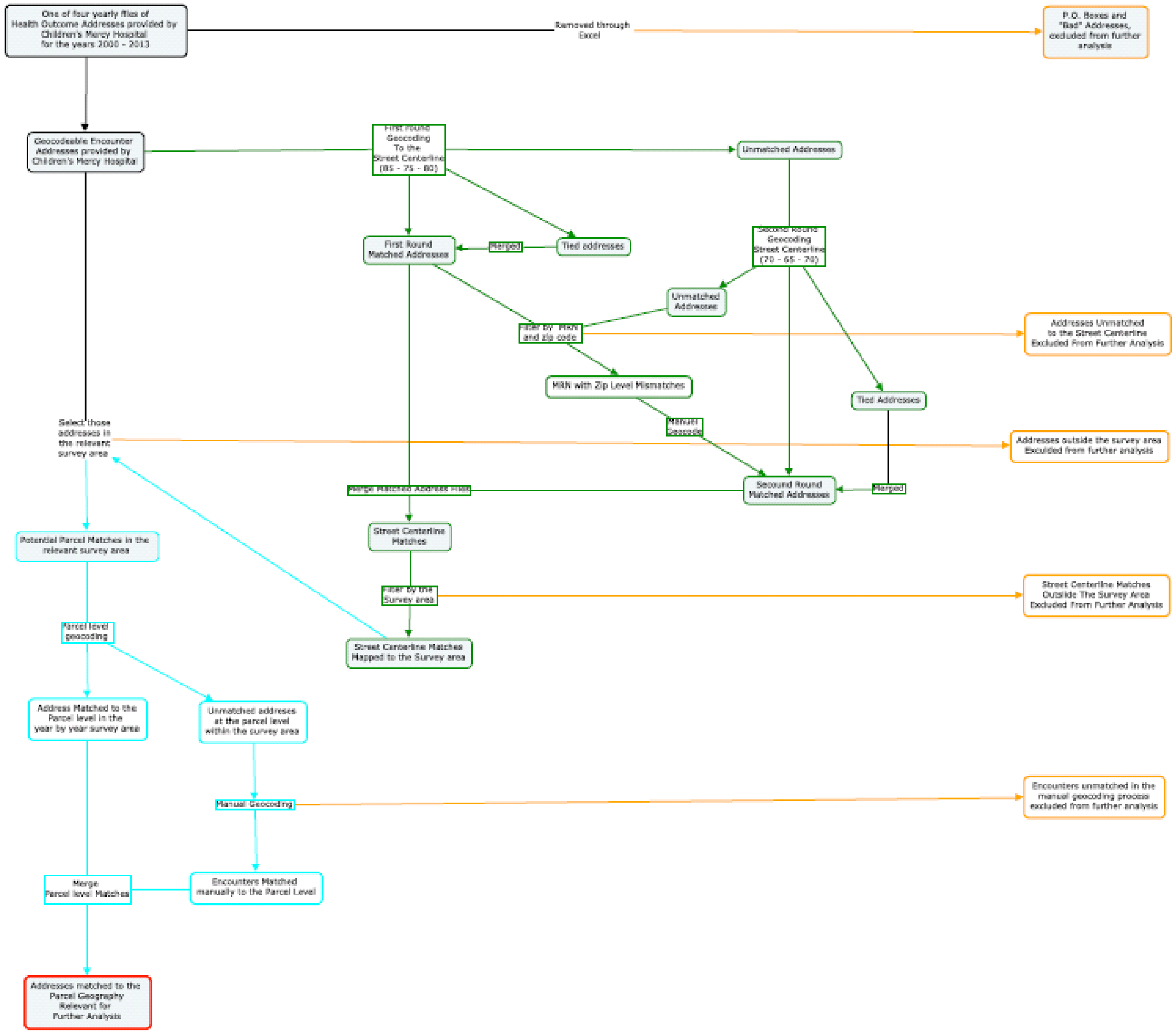


Figure 5: Diagram of the iterative address geocoding process

Step 1: First filtration process

First we remove those encounters with addresses that cannot be matched to a street address. Some addresses were listed explicitly as “bad” by the hospital others were specified as a P.O. Box and still others indicated a geography, such as an address in another state, inconsistent with our study. This sort of first step filtration is standard practice in parcel geocoding. (Sonderman et al 2012) These unmatchable encounters are not considered relevant to any future match rates. Though by sharing the characteristic of being un-geocode-able these encounters constitute a subset of encounters that can be analyzed in their own right.

Step 2: First Street Centerline Geocoding

Using a centerline address locator derived from the MARC street centerline file those health encounters with complete addresses are geocoded to create the first point layer shapefile. The criteria data for this geocoding operation are street address (number, direction, name and street type) and zip code. For this first round operation the Geocoding Options with regard to Spelling sensitivity, Minimum candidate score and Minimum match score are left at their default values. The address match process creates three classes of health encounters in ArcMap. Matched encounters, those the address locator has been able to place along the street centerline shape file. Tied encounters, those the address locator has found more than one possible match to the street centerline geography. Unmatched encounters, those unmatched to the street centerline geography. Because this first stage of geocoding is designed principally to identify likely matches for future parcel matches, the tied encounters are grouped in with the matched encounters while unmatched encounters were passed along to the next iteration of the geocoding process.

Step 3: Second Street Centerline Geocoding

Using the same MARC-derived address locator, the geocoding program is run a second time but only on those encounters unmatched in step 2. In this round of geocoding the Geocoding Options are altered such that Spelling Sensitivity, Minimum candidate score and Minimum match score are lower and thus more likely to produce an automated match when faced with an ambiguous address. Again, this process creates three classes of health encounters, and tied encounters are grouped in with the matched encounters. This time a filtering process is added to check for miss-specified zip codes among the still unmatched encounter addresses. The filtering is done by an automated linking of matched and unmatched encounters by MRN followed by the manual comparison and correction of miss-specified zip-codes among the unmatched encounters. Those encounters with addresses either matched or tied in step 3 are now merged with those same health encounter classes from step 2. Those that remain unmatched are, similar to the un-geocode-able addresses from step 1, set aside for future non-spatial analysis.

Step 4: Neighborhood selection and parcel level Geocoding

Steps 1 thru 3 creates a point level shapefile of health encounters for each year of the study. Step 4 selects those matched/tied encounters from this shapefile that in proximity to the target NHCS geography then matches these selected encounters to the surveyed parcels. The selection is done by generating a polygon geography representing the general extent of each NHCS (or creating a buffer around the parcels themselves) and using it to select those points likely to match the parcels in question. Next, a parcel address locator is derived from the target NHCS using the criteria fields of Street Address and Zip code. Recall that each NHCS geography is different and some years saw several housing condition surveys. Running the NHCS parcel level address locator on the addresses selected above produces the now familiar classes of matched, tied and unmatched.

A comparison of this result with the results of the parcel match that would occur using the NHCS address locator on the health encounter data without first steps 1 through 3 is instructive. (This comparison should be performed for quality assurance reasons to ensure there are no encounters matched to parcels that are not matched to centerlines and vice versa) Each geocoding operation returns to the original address data – there should be no difference among the matched/tied classes. The difference between the unmatched classes of running and not running steps 1 through 3 is the object of our process. These steps result in an order of magnitude fewer unmatched encounters than we would otherwise be facing. This performs two functions. First, it allows the calculation of a meaningful match rate relating health encounters to parcels with observed housing conditions. Second it streamlines the process of identifying unmatched records that warrant manual re-matching.

Step 5: Manual rematch process

We now have two point shapefiles derived from the same health encounter data. One is matched to the street centerline geography, it is set aside for future analysis. The task now is to examine that portion of the original health data that has not been matched to the parcel level and is likely to be matched through a manual inspection of the encounter address and the potentially matching parcel address. First, sort the unmatched parcel matches by a common property such as zip code or neighborhood. This sorting concentrates our focus on one neighborhood at a time, saving the effort of scanning between disparate sections of the city with each new record. Next we examine each health encounter address for accuracy in spelling, nomenclature and completeness. A reason for no match might be as simple as a misspelling of one of the address components, missing directional data or an incorrect zip code. Then we examine the target NHCS parcel geography for addresses that are not indicated in the address locator. Address locators derived from a parcel geography assign a single address to each polygon feature. Kansas City, like many cities, has types of household structures, including multifamily, duplexes, and apartment buildings, where a single parcel is the location for many addresses. A related problem is a mismatch between the postal address used by the household and the administrative address used by the NHCS.

The method for manually comparing unmatched health encounter addresses with potential parcel matches warrants explicit discussion. In this process secondary sources of information are used to locate the potential parcel match associated with the unmatched encounter addresses (Goldberg 2008). To get a general idea of where in the city the unmatched address is located, it is queried using an online resource such as google maps or the Kansas City Parcel Viewer. Visual images in google maps are helpful for finding in-fill addresses (such as 2010½ or 2010c). After identifying the location of the potential parcel match its corresponding NHCS parcel record is consulted for any reference to the unmatched address. The 'legal' attribute is critical in this search as it can contain reference to a range of addresses associated with the identified parcel. This is process is required for every apartment building, complex and multi-family dwelling in the survey area but the rematch process is not limited to these cases.

After examining every unmatched address in proximity to the NHCS parcels perform the same inspection process on the class of tied encounters to clarify the location of the otherwise ambiguous address. Once the list of unmatched and tied encounters is examined they are merged with those automatically matched from step 4 into a single file. This produces a single point shape file for health encounters from the year in question associated with the parcel geography of a single NHCS.

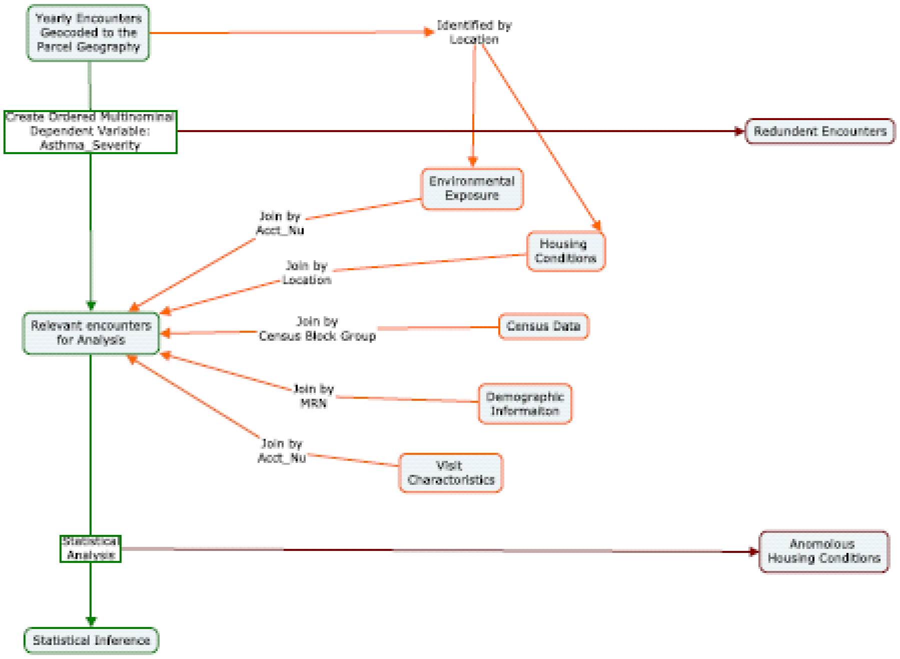
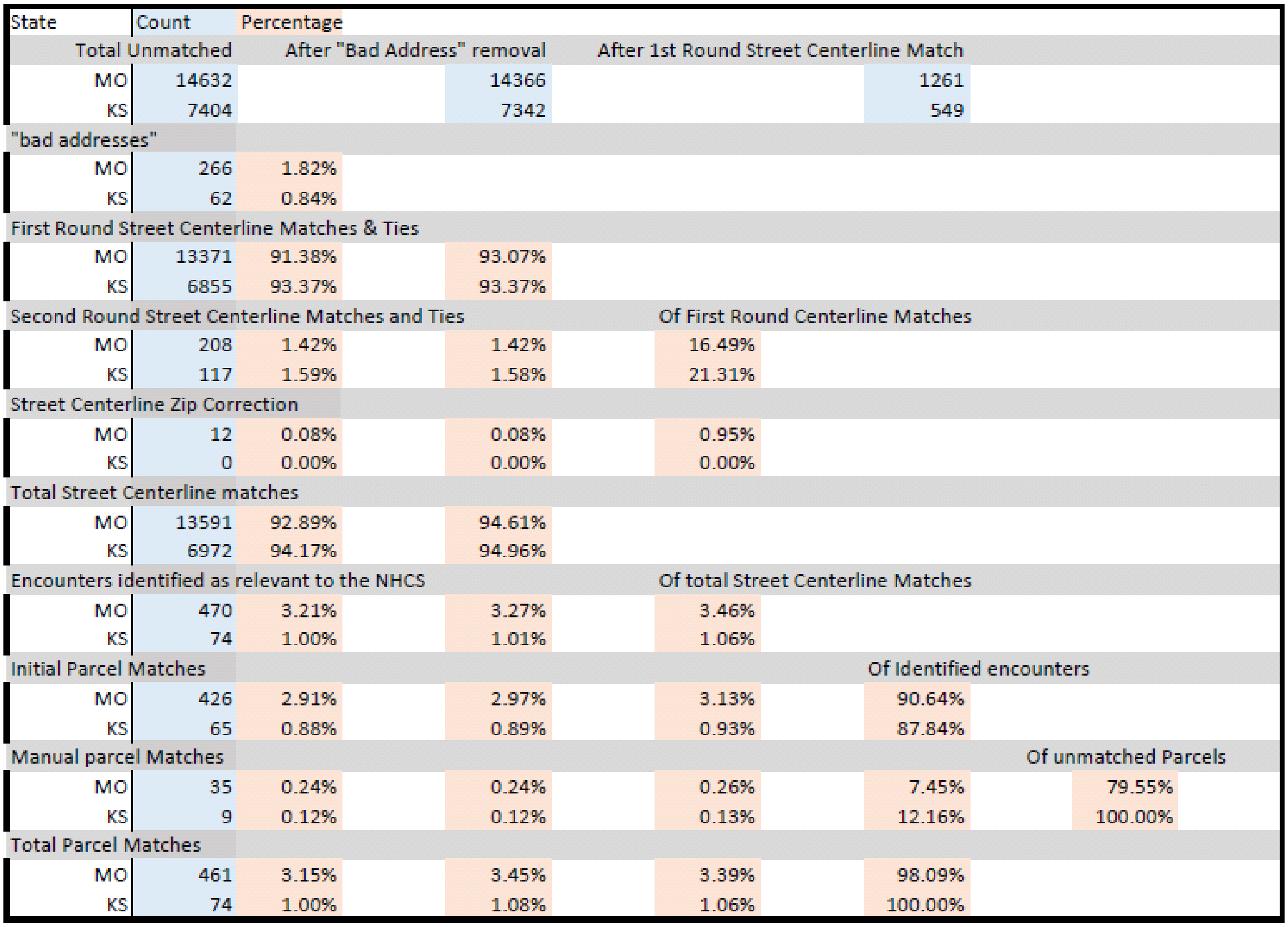


Figure 6 Data assembly process for Parcel level matches.

Step 6

Repeat steps 1 through 5 for each year of the study and for each of the NHCS. This will produce a point shapefiles geocoded to the street-centerline geography for each year of the study and a point shapefile for each of the NHCS geocoded to the parcel geography. Now, merge all the street centerline encounters into a single shapefile. Do the same with the parcel level encounters. Each of these geography specific shapefiles are then available for the data assembly process whereby patient specific information can be associated with housing conditions, visit characteristics, demographic information, environmental and block group census data via the unique identifiers. Figure 6 is a diagram of the data assembly process. At the end of step 6 the data is ready for statistical analysis.



| State | Count | Percentage | | | | |
|---|---|---|---|---|---|---|
| Total Unmatched | | | After "Bad Address" removal | | After 1st Round Street Centerline Match | |
| MO | 14632 | | 14366 | | 1261 | |
| KS | 7404 | | 7342 | | 549 | |
| "bad addresses" | | | | | | |
| MO | 266 | 1.82% | | | | |
| KS | 62 | 0.84% | | | | |
| First Round Street Centerline Matches & Ties | | | | | | |
| MO | 13371 | 91.38% | 93.07% | | | |
| KS | 6855 | 93.37% | 93.37% | | | |
| Second Round Street Centerline Matches and Ties | | | | Of First Round Centerline Matches | | |
| MO | 208 | 1.42% | 1.42% | 16.49% | | |
| KS | 117 | 1.59% | 1.58% | 21.31% | | |
| Street Centerline Zip Correction | | | | | | |
| MO | 12 | 0.08% | 0.08% | 0.95% | | |
| KS | 0 | 0.00% | 0.00% | 0.00% | | |
| Total Street Centerline matches | | | | | | |
| MO | 13591 | 92.89% | 94.61% | | | |
| KS | 6972 | 94.17% | 94.96% | | | |
| Encounters identified as relevant to the NHCS | | | | Of total Street Centerline Matches | | |
| MO | 470 | 3.21% | 3.27% | 3.46% | | |
| KS | 74 | 1.00% | 1.01% | 1.06% | | |
| Initial Parcel Matches | | | | | Of Identified encounters | |
| MO | 426 | 2.91% | 2.97% | 3.13% | 90.64% | |
| KS | 65 | 0.88% | 0.89% | 0.93% | 87.84% | |
| Manual parcel Matches | | | | | | Of unmatched Parcels |
| MO | 35 | 0.24% | 0.24% | 0.26% | 7.45% | 79.55% |
| KS | 9 | 0.12% | 0.12% | 0.13% | 12.16% | 100.00% |
| Total Parcel Matches | | | | | | |
| MO | 461 | 3.15% | 3.45% | 3.39% | 98.09% | |
| KS | 74 | 1.00% | 1.08% | 1.06% | 100.00% | |

Figure 7: 2012 match rates for each step of the iterative process.

## 4 Results:

This section discusses that process with regard to the identified criteria of completeness, positional accuracy, repeatability, and cost. These issues are addressed one by one.

Completeness:

The method we outline is designed to allow us to determine a warranted match rate for the parcel geography and thereby determine the completeness of our results. For this discussion we turn to figure 7, the match rates and numerical counts for each step of the iterative process applied to the year 2012 data. The figure begins at the top left with a record of the total number of unmatched Asthma encounters provided by Children's Mercy Hospital. Moving down this figure tracks the individual steps in the iterative match process. Moving to the right in the figure displays the match rate of each step with reference to the several that came before.

The raw match rate, the number of parcel level matches divided by the number of initial health encounters, is 3.5% for encounters in Missouri and 1% for those in Kansas. However, this raw rate is inadequate because it considers all encounters for the year, those with "bad" addresses and all, and does not focus account those encounters that can reasonably be identified as possible matches. Focusing on these warranted encounters, those identified through our iterative approach, indicated that reasonable match rates after manual geocoding at the parcel level to be 98% for encounters in Missouri and 100% of those in Kansas. Reference to figure 7 also indicates the completeness of the geocoding to the street centerline geography. In the case of this geography the entire range of data are likely matches and by using the centerline address locator we observe match rates of over 94% for both Missouri and Kansas addresses. Appendix 2, Parcel Match Rate for All Years, details the match rates for all years (2000 – 2012) in which a NHCS was completed. The centerline match rate for the complete data set is 91.1%. Yearly centerline match rates range from 87.9% in the year 2000 to 94.3% in 2011. The Parcel Match for the complete data set is 60.4% with yearly match rates ranging from 35.2% in 2005 to 100% in 2011. Causes for this discrepancy in match rates from year to year is discussed below in the section on repeatability.

Positional accuracy:

Our iterative process is in keeping with the best practices for positional accuracy in geocoding (McElroy et al 2003; Rushton et al 2006; Manson et al 2008; Jacquez 2012; Sonderman et al 2012). Aim 1 of KC-Heart, examining the relationship with exterior housing conditions and the intensity of childhood asthma, is dependent on a spatial merger of points and polygons in GIS. A point layer can be easily associated with a polygon layer if they overlap in the conceptual space of the GIS. The required overlap was generated by matching to the centroid of the parcel based using an address locator based on the target geography. The veracity of the point position is confirmed in the spatial association of the health encounter with the housing conditions parcel. The positional accuracy of all health encounters is confirmed via a complete merge of parcel data – no encounters were placed such that a point did not have a polygon match. This bodes well for the statistical legitimacy of our study.

Repeatability:

The health data we received from CMH and KCHD are subset by year, by state and by health outcome. These divisions in the data are maintained because of limitations in computing power (ESRI's ArcMap is a very efficient program, still the undivided data taxed that capabilities of our resources) and because these divisions echo the structure of the NHCS that guide the parcel geocoding process. These divisions of the data entail repeating the geocoding process for each data subset, while internal to the iterative process are several distinct geocoding processes. For the health data from multiple geocoding operations to be joined into a single file for statistical analysis the operations must be methodologically consistent. With several GIS analysts working on different stages of the geocoding simultaneously a clear method and consistent set of instructions to guide the process when minute variations and anomalies in the source and target data are discovered. For this reason precise geocoding instructions, of which appendix 1 is an example, were generated for each subset of data. Thus following the documented instructions is required for the repeatability of the overall process.

Repeatability of the process is challenged where ever a potential match is left to the judgment of the GIS analyst or is dependent on data not included in the project geography of KC-heart. An important instance where personal judgement could enter into the iterative process is in the manual rematch step. To limit the discretion of the analyst and increase the likelihood that our geocoding results could be reconstructed in their entirety we required any manual match to be based on data internal to the project. This decision is guided by the logic of HIPPA, which dictates in effect that the health encounter data will never be shared in its raw form with anyone outside of our research consortium. Google maps, street-view, and local parcel viewers can suggest a matching location but if the NHCS parcel file contains no reference to the address in question no match would take place.

There is a tradeoff here that favors the internal replication of results using the NHCS data and project geography over completeness of the parcel level matches. The most common issue to spur the manual rematch process are several address assigned to the same parcel, recall the address locator uses a single address to represent each parcel. In this case the "legal" field of the multi-household polygon generally contains reference to the range of subordinate addresses associated with that location. However, if the "legal" field is missing, limited or corrupt the manual match rate suffers. This is the case with several years of the NHCS and we observe the parcel match rate suffers for it. The decision was made to forego the potential mismatches and unrepeatability which would follow opening up the manual rematch process to the whimsy of the analyst for the precision and the repeatability of the process described.

Cost:

In an ideal world the cost of geocoding in health research would be irrelevant to the consideration of research method. However, health researchers are under constant pressure to cut costs and stretch existing funding. Parcel geocoding, despite being preferred for analytical purposes, because it entails a manual step, with its time consuming address by address inspection, can be prohibitively expensive. This was the case for KC-Heart; the target parcel geography, corresponding to a sometimes quite small subset of the total municipal geography, entails a large volume of encounters to be investigated, one at a time, via manual inspection process, despite the majority of these encounters being located at a remove from the target geography. Instead of embarking on the costly investigation of every parcel, specific geographies were constructed to identify the warranted set of potential matches for each NHCS. This process not only generated a relevant measure for match rate, as described above, it also limited the manual matching process to those health encounters which are likely to correspond to parcels with housing conditions data. Appendix 2, Parcel Match Rate for all years, lists the number of parcels excluded from manual inspection for each year of the NHCS. Then, using baseline rate of 50 addresses examined per hour derived from internal record keeping and the geocoding literature (McElroy et al 2002, Goldberg et al 2008), we estimate our iterative match process to have saved approximately 3500 hours of work (and the significant labor cost associated with that work) necessary to arrive at similar results using the standard method.

## Discussion:

Mapping, data storage and spatial analysis are increasingly important tools in epidemiological research and public health practice, prevention, and policy (Edwards et al. 2013). Developments in GIS including computing power, intuitive user interfaces, the proliferation of available data, and computational ability move in consort with its adoption as a standard research tool (Allen and Coffey 2011). GIS is now regularly applied to address research questions in epidemiology and public health, urban planning, environmental science, law enforcement, forestry, fire prevention, and economic analysis. (Chetty et al 2016; Hernandez 2009; Morenoff 2003; Richardson 2010, Chuvieco and Salas 1996) The general usefulness of GIS in disparate fields make it an ideal methodological tool for interdisciplinary research.

However, the strength of these tools requires that the goal is always to produce complete and accurate geocoding results. One of the issues addressed above is that quantifying such criteria (completeness and accuracy) with regard to an inconsistent parcel geography is expensive and time consuming. This paper breaks new ground by developing a warranted match rate where one is not immediately obvious. At the same time, we document that our method produces a diverse collection of disaggregated spatio-temporal shapefiles that avail themselves to a variety of research inquiries capable of addressing a number of prominent concerns in this research, such as the ecological fallacy, MAUP and UGCoP. Thus, by contributing to existing literature of best practices for geocoding, we argue it is critical to continue to expand and support interdisciplinary efforts in public and preventative health research.

To this end, the initial analysis of the assembled data from the year 2001 indicate a weak but statistically present influence of housing conditions on the severity of the asthma encounter[5]. These findings recommend expanding the statistical analysis to include the full range of health encounter data. As indicated above the legitimacy of these statistical results depend on the veracity of the address match process. Legitimate modeling of the relationship between health and housing is not possible if the health records are not matched to the proper home address.

## References

Abellan, TJ., Richardson, S. and Best, N. (2008) Use of Space-Time Models to Investigate the Stability of Patterns of Disease. *Environmental Health Perspectives*. 116 (8): 1111-1119.

Allen, D.W. (2009) *GIS Tutorial: Spatial Analysis Workbook*. Redlands, CA. ESRI Press.

Allen, D.W., Coffey, J.M. (2011) *GIS Tutorial: Advanced Workbook*. Redlands, CA. ESRI Press.

Briant, A., Combes, P.P., Lafourcade, M. (2010) "Dots to boxes: Do the size and shape of spatial units jeopardize economic geography estimations?" *Journal of Urban Economics*. Vol. 67, 287-302.

Brauer, M. Hoek, G. Van Vliet, P Meliefste, K. Fischer, P. Wijga, Koopman, L.P. Neijens, H.J. Ger ritsen, J. Kerkhof, M. and J. Heinrich (2002) "Air pollution from traffic and the development of

[5] For a preliminary discussion of our findings relating Asthma severity and housing conditions please see WP 1602-01

respiratory infections and asthmatic and allergic symptoms in children". American journal of respiratory infections and asthmatic and allergic symptoms in children. 166, no. 8: 1092 – 1098.

Chetty, R., Hendren, N., Lin, F., Majerovitz, J., Scuderi, B. (2016) Childhood Environment and Gender Gaps in Adulthood. NBER Working Paper Series. Working Paper 21936.

Chetty, R., Hendren, N., Kline, P., Saez, E. (2014) Where is the Land of Opportunity? The Geography of Intergenerational Mobility in the United States. *Quarterly Journal of Economics.* 129(4): 2633-2679.

Chuvieco, E. and J. Salas (1996) "Mapping the spatial distribution of forest fire danger using GIS" *International Journal of Geographical Information Systems* 10 (3): 333 -345

Coleman-Jensen, A., Nord, M., Andrews, M., Carlson, S. (2011) "Household Food Security in the United States in 2010". United States Department of Agriculture. Economic Research Service. Economic Research Report Number 125.

Dearwent, S.M. Jacobs, R.R. and J.B. Halbert (2001) "Locational uncertainty in georeferencing public health datasets". *Journal of exposure analysis and environmental epidemiology.* 11: 329 - 334

Edwards, SE., Strauss, B. and Miranda, ML. (2014) Geocoding Large Population-level Administrative Datasets at Highly Resolved Spatial Scales. *Transactions in GIS*. 18 (4): 586-603.

Goldberg, D.W., Wilson, J.P., Knoblock, C.A., BRitz, B., and M.G. Cockburn. (2008). "An effective and efficient approach for manually improving geocoded data." *International Journal of Health Geographics.*7:60

Humphries, A., Wilson, B., Reddy, M., Shroba, J. and Ciaccio, C. (2015) An Association Between Pediatric Food Allergy and Food Deserts. *Journal of Allergy and Clinical Immunology*. 135 (2); AB255.

Jacquez, G. (2012) "A research agenda: Does geocoding positional error matter in health." *Spatial and Spatio-temoral Epidemiology*. 3: 7-16.

Johnston, R. (2005) Geography and GIS in Paul Longley et al. eds. *Geographical Information Systems: Principles, Techniques, Management, and Applications*. John Wiley & Sons, Inc. 27-35.

Kwan, Mei-Po (2012) "The Uncertain Geographic Context Problem". *Annals of the Association of American Geographers.* 102(5) : 958 – 968.

Logan, R., Zhang, W., Xu, H. (2010) "Applying Spatial Thinking in Social Science Research". GeoJournal. 75:15-27.

Manson S. M. Sander, H. A. Ghosh, D. Oakes, J. M. Orfield Jr, M. W. Craig, W. J. Luce, T.F. Myott, E. and S. Sun (2009) “Parcel data for research and policy”. *Geography Compass*. 3/2, 698 - 726

Matsui, E.C., Abramson, S.L., Sandel, M.T. (2016) “Indoor Environmental Control Practices and Asthma Management”. *Pediatrics*. 138(5): e1-e11.

McElroy, J.A., Remington, P.L., Trentham-Dietz, A., Robert, S.A., Newcomb, P.A. (2003) “Geocoding Addresses from a Large Population-Based Study: Lessons Learned”. *Epidemiology*. 14(4) 399-407

Messner, S., Anselin, L., Baller, D., Hawkins, D., Deane, G., Tolnay, S. (1999) “The Spatial Patterning of County Homicide Rates: An Application of Exploratory Data Analysis”. *Journal of Quantitative Criminology*. 15( 4): 423-450.

Miranda, M.L. and Edwards, S.E. (2011) “Use of Spatial Analysis to Support Environmental Health Research and Practice”. *N.C. Medical Journal*. 72(2): 123-135.

Monmonier, M. (1996) *How to Lie with Maps: Second Edition*. University of Chicago Press.

Morenoff, J. (2003) Neighborhood Mechanisms and the Spatial Dynamics of Birth Weight. *American Journal of Sociology* 108: 976-1017.

Openshaw, S. (1984) “The modifiable areal unit problem”, *Concepts and Techniques in Modern Geography*, 38:41.

Rauh, V., Landrigan, P., Claudio, L. (2008). Housing and Health Intersection of Poverty and Environ mental Exposures. *Annuals of the New York Academy of Science*. 1136: 276-288.

Rabito, F.A., Iqbal, C.F. Shorter, P. Osman, P.E. Philips, E. Langlois, and L.E. White. (2007) “The association between demolition activity and children’s blood lead levels” *Environmental research* 103, no. 3: 345-351.

Richardson, K. (2010) “Exploring Food Environments: Assessing Access to Nutritious Food” *ArcUser*. Fall, 50-52.

Rushton, G., Armstrong, M., Gittler, J., Greene, B.R., Pavilik, C.E., West, MM., and Zimmerman, DL. (2006) “Geocoding in Cancer Research”. *American Journal of Preventative Medicine*. 30 (2S): S16-S24.

Sonderman, J.S., Mumma, M.T., Cohen, S.S., Cope, E.L., Blot, W.J. Signorello, L.B. “A multi-stage approach to maximizing geocoding success in a large population-based cohort study through automated and interactive processes”. *Geospatial Health*. 6(2): 273-284

Schuch, L. Curtis, A and J. Davidson (2016) “Reducing lead exposure risk to vulnerable populations: A proactive geographic solution”. *Annals of the American Geographer*. 0(0): 1 - 19

Stokol, D. (1996) "Translating Social Ecological Theory into Guidelines for Community Health Promotion". *American Journal of Public Health Promotion*. 67:1, 282-298.

Susser, M. (1994) "The Logic in Ecological: I. The Logic of Analysis". *American Journal of Public Health*. 84:5, 825-830.

Whitsel, E.A. Rose, K.M. Wood, J.L. Henley, A.C. Liao, D. and G. Heiss. "Accuracy and Repeatability of Commercial Geocoding" *American Journal of Epidemiology.* 160(10): 1023-1029

Zandbegen, P (2007) "Influence of geocoding quality on environmental exposure assessment of children living near high traffic roads" *BMC Public Health*. 7 (37)

Zandbergen, P (2008) "Geocoding quality and implications for spatial analysis". *Geography Compass* 3/2, 547-680.

# Appendix 1: Geocoding Process for the 2012 Well Child Encounters

## Map Data

Survey Parcels -> SurveyParcels2012 : \\134.193.143.54\sfiles\GeoData\NHCS\NHCS_Geo\2012

Survey Neighborhood -> none in evidence

Marc Streets -> Marc_Streets : \\134.193.143.54\sfiles\GeoData\ProjectGeography\MARC_clns2015\marc_streets.gdb

Missouri Zip Codes -> MOzipcodes \\134.193.143.54\sfiles\GeoData\NHCS\NHCS_Geo\2012

Children's Mercy Data -> 'Mo Controls$' : \\134.193.143.54\sfiles\DataSetsI\Asthma_Controls\2012\Cleaned Addresses.xlsx

SSL 2012 Missouri Asthma Control Encounters ->MoC_SSL_2012 : \\134.193.143.54\sfiles\GeoData\NHCS\NHCS_Geo\2012

Housing Survey Data -> see note below

Updated Addresses -> none available

Address Locator -> SurveyParcels2012 : \\134.193.143.54\sfiles\GeoData\NHCS\NHCS_Geo\2012

*note: There is no survey area available in the 2012 folder. The Survey data has been appended to the survey parcel layer in the 2012 Asthma diagnosis parcel level geocoding process. The Address locator was constructed in that same process.

## Initial Summary Statistics for GeoMo2012

Surveyed Parcels: 6094

Asthma Control Encounters for MoC_SSL_2012: 20631

Encounters Matched to the Street Center line: 17875

## Process for Geocoding to Surveyed Parcels:

1. Add the above data to a new ArcMAP. Name the file MoControl2012 and save it in the 2012 Folder
2. Confirm: NAD_1983_StatePlane_Missouri_West_FIPS_2403_Feet
   And Projection: Transverse_Mercator
3. Trim the Missouri zip codes and the MARC street maps to the area of the surveyed parcels.
4. Select by location those asthma control encounters contained in the file MoC_SSL_2012. Select features that are within a distance of 100 feet of the surveyParcels2012zip layer. Chose to create a new layer from the selected features.
   N=1217
5. Using the SurveyParcels2012 address locator geocode 'Mo Controls$' to the parcel geography. Name the new layer CMH_CParcels_2012 and save in the TroubleShooting folder.

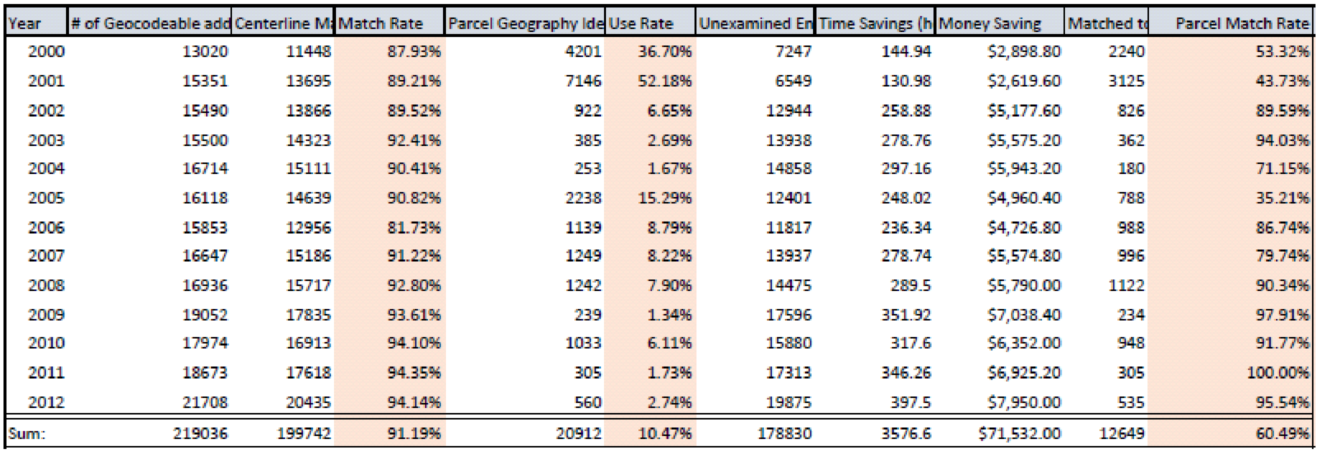
N=1080

6. By inspection we see that the number of unmatched asthma encounters = 1217-1080 = 137.

Manual Geocoding Process

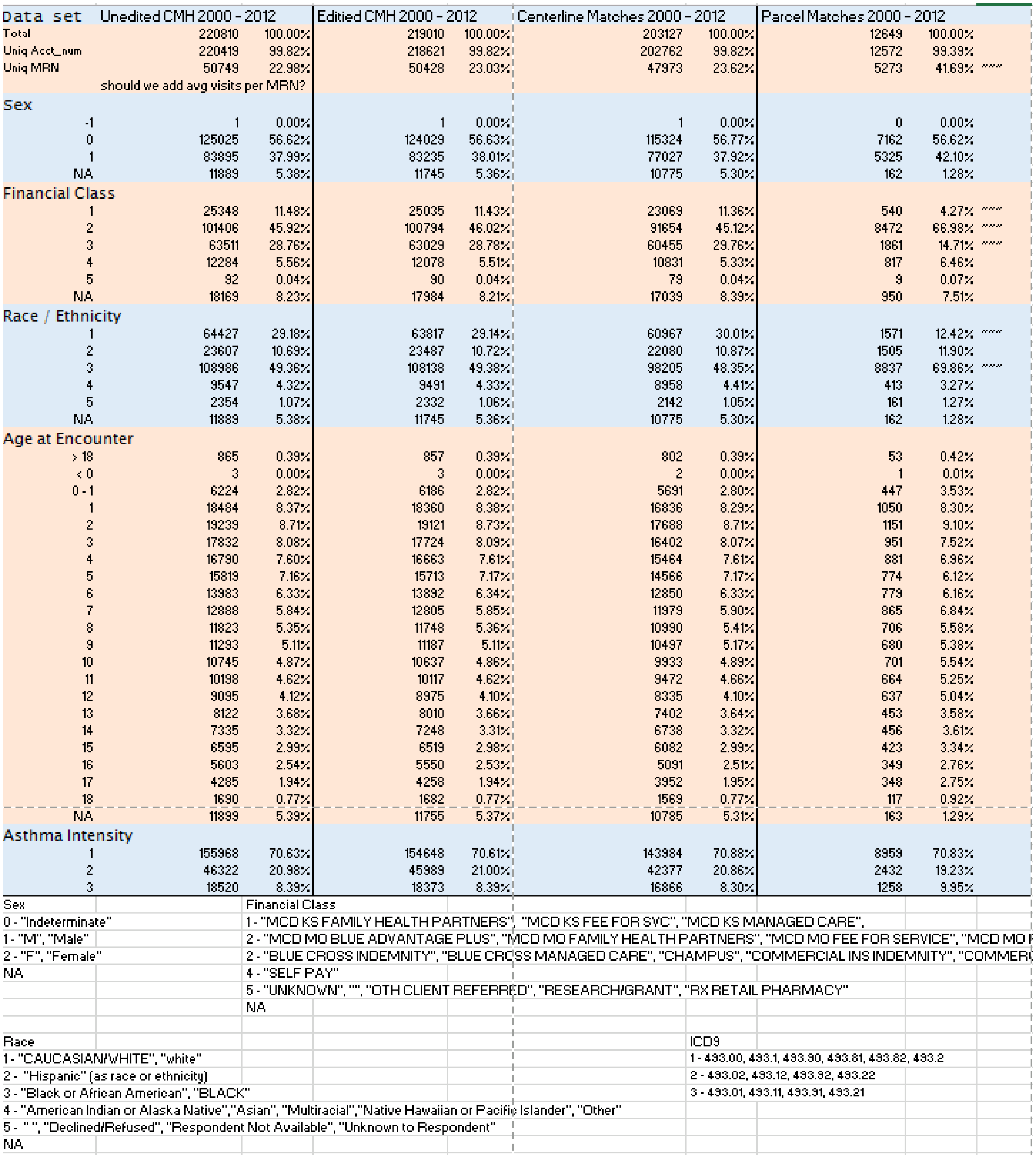
7. Export the data in MoC_SSL_2012 selected features. Save the feature class as MoC_SSL_SA_2012 in the Troubleshooting folder
8. Join the MoC_SSL_SA_2012 layer to the CMH_CParcels_2012 layer by Account_Nu keeping only the matching records
   N=1217
9. By inspection we observe that 1077 of the 1080 parcel matches remain. To find these three missing parcels remove the join from #8. Join MoC_SSL_2012 to CMH_CParcels_2012 by Account_Nu. Now select by Attributes “"MoC_SSL_2012.Status" = 'U' AND "CMH_CParcels_2012.Status" = 'M' or "CMH_CParcels_2012.Status" = 'T'”
10. Export The selected records from #9 to a dBase file. Save as MoC2012MopUp.1 in the troubleshooting.
11. Remove Join from #10 and repeat step #8. Select “"CMH_CParcels_2012.Status" = 'U'”. Export the selected records as a dBASE table. Save as MoC_2012_MopUp in the TroubleShooting folder.
12. Delete the redundant and extraneous fields from the MopUp files.
    N=140

## Appendix 2: Parcel Match Rate for All Years



| Year | # of Geocodeable add | Centerline M | Match Rate | Parcel Geography Ide | Use Rate | Unexamined En | Time Savings (h | Money Saving | Matched t | Parcel Match Rate |
|---|---|---|---|---|---|---|---|---|---|---|
| 2000 | 13020 | 11448 | 87.93% | 4201 | 36.70% | 7247 | 144.94 | $2,898.80 | 2240 | 53.32% |
| 2001 | 15351 | 13695 | 89.21% | 7146 | 52.18% | 6549 | 130.98 | $2,619.60 | 3125 | 43.73% |
| 2002 | 15490 | 13866 | 89.52% | 922 | 6.65% | 12944 | 258.88 | $5,177.60 | 826 | 89.59% |
| 2003 | 15500 | 14323 | 92.41% | 385 | 2.69% | 13938 | 278.76 | $5,575.20 | 362 | 94.03% |
| 2004 | 16714 | 15111 | 90.41% | 253 | 1.67% | 14858 | 297.16 | $5,943.20 | 180 | 71.15% |
| 2005 | 16118 | 14639 | 90.82% | 2238 | 15.29% | 12401 | 248.02 | $4,960.40 | 788 | 35.21% |
| 2006 | 15853 | 12956 | 81.73% | 1139 | 8.79% | 11817 | 236.34 | $4,726.80 | 988 | 86.74% |
| 2007 | 16647 | 15186 | 91.22% | 1249 | 8.22% | 13937 | 278.74 | $5,574.80 | 996 | 79.74% |
| 2008 | 16936 | 15717 | 92.80% | 1242 | 7.90% | 14475 | 289.5 | $5,790.00 | 1122 | 90.34% |
| 2009 | 19052 | 17835 | 93.61% | 239 | 1.34% | 17596 | 351.92 | $7,038.40 | 234 | 97.91% |
| 2010 | 17974 | 16913 | 94.10% | 1033 | 6.11% | 15880 | 317.6 | $6,352.00 | 948 | 91.77% |
| 2011 | 18673 | 17618 | 94.35% | 305 | 1.73% | 17313 | 346.26 | $6,925.20 | 305 | 100.00% |
| 2012 | 21708 | 20435 | 94.14% | 560 | 2.74% | 19875 | 397.5 | $7,950.00 | 535 | 95.54% |
| Sum: | 219036 | 199742 | 91.19% | 20912 | 10.47% | 178830 | 3576.6 | $71,532.00 | 12649 | 60.49% |

## Appendix 3: Summary Stats for Unedited, Edited, Centerline and Parcel Level Health Encounters



| Data set | Unedited CMH 2000 - 2012 | | Editied CMH 2000 - 2012 | | Centerline Matches 2000 - 2012 | | Parcel Matches 2000 - 2012 | | |
|---|---|---|---|---|---|---|---|---|---|
| Total | 220810 | 100.00% | 219010 | 100.00% | 203127 | 100.00% | 12649 | 100.00% | |
| Uniq Acct_num | 220419 | 99.82% | 218621 | 99.82% | 202762 | 99.82% | 12572 | 99.39% | |
| Uniq MRN | 50749 | 22.98% | 50428 | 23.03% | 47973 | 23.62% | 5273 | 41.69% | ~~~ |
| | should we add avg visits per MRN? | | | | | | | | |
| Sex | | | | | | | | | |
| -1 | 1 | 0.00% | 1 | 0.00% | 1 | 0.00% | 0 | 0.00% | |
| 0 | 125025 | 56.62% | 124029 | 56.63% | 115324 | 56.77% | 7162 | 56.62% | |
| 1 | 83895 | 37.99% | 83235 | 38.01% | 77027 | 37.92% | 5325 | 42.10% | |
| NA | 11889 | 5.38% | 11745 | 5.36% | 10775 | 5.30% | 162 | 1.28% | |
| Financial Class | | | | | | | | | |
| 1 | 25348 | 11.48% | 25035 | 11.43% | 23069 | 11.36% | 540 | 4.27% | ~~~ |
| 2 | 101406 | 45.92% | 100794 | 46.02% | 91654 | 45.12% | 8472 | 66.98% | ~~~ |
| 3 | 63511 | 28.76% | 63029 | 28.78% | 60455 | 29.76% | 1861 | 14.71% | ~~~ |
| 4 | 12284 | 5.56% | 12078 | 5.51% | 10831 | 5.33% | 817 | 6.46% | |
| 5 | 92 | 0.04% | 90 | 0.04% | 79 | 0.04% | 9 | 0.07% | |
| NA | 18169 | 8.23% | 17984 | 8.21% | 17039 | 8.39% | 950 | 7.51% | |
| Race / Ethnicity | | | | | | | | | |
| 1 | 64427 | 29.18% | 63817 | 29.14% | 60967 | 30.01% | 1571 | 12.42% | ~~~ |
| 2 | 23607 | 10.69% | 23487 | 10.72% | 22080 | 10.87% | 1505 | 11.90% | |
| 3 | 108986 | 49.36% | 108138 | 49.38% | 98205 | 48.35% | 8837 | 69.86% | ~~~ |
| 4 | 9547 | 4.32% | 9491 | 4.33% | 8958 | 4.41% | 413 | 3.27% | |
| 5 | 2354 | 1.07% | 2332 | 1.06% | 2142 | 1.05% | 161 | 1.27% | |
| NA | 11889 | 5.38% | 11745 | 5.36% | 10775 | 5.30% | 162 | 1.28% | |
| Age at Encounter | | | | | | | | | |
| > 18 | 865 | 0.39% | 857 | 0.39% | 802 | 0.39% | 53 | 0.42% | |
| < 0 | 3 | 0.00% | 3 | 0.00% | 2 | 0.00% | 1 | 0.01% | |
| 0 - 1 | 6224 | 2.82% | 6186 | 2.82% | 5691 | 2.80% | 447 | 3.53% | |
| 1 | 18484 | 8.37% | 18360 | 8.38% | 16836 | 8.29% | 1050 | 8.30% | |
| 2 | 19239 | 8.71% | 19121 | 8.73% | 17688 | 8.71% | 1151 | 9.10% | |
| 3 | 17832 | 8.08% | 17724 | 8.09% | 16402 | 8.07% | 951 | 7.52% | |
| 4 | 16790 | 7.60% | 16663 | 7.61% | 15464 | 7.61% | 881 | 6.96% | |
| 5 | 15819 | 7.16% | 15713 | 7.17% | 14566 | 7.17% | 774 | 6.12% | |
| 6 | 13983 | 6.33% | 13892 | 6.34% | 12850 | 6.33% | 779 | 6.16% | |
| 7 | 12888 | 5.84% | 12805 | 5.85% | 11979 | 5.90% | 865 | 6.84% | |
| 8 | 11823 | 5.35% | 11748 | 5.36% | 10990 | 5.41% | 706 | 5.58% | |
| 9 | 11293 | 5.11% | 11187 | 5.11% | 10497 | 5.17% | 680 | 5.38% | |
| 10 | 10745 | 4.87% | 10637 | 4.86% | 9933 | 4.89% | 701 | 5.54% | |
| 11 | 10198 | 4.62% | 10117 | 4.62% | 9472 | 4.66% | 664 | 5.25% | |
| 12 | 9095 | 4.12% | 8975 | 4.10% | 8335 | 4.10% | 637 | 5.04% | |
| 13 | 8122 | 3.68% | 8010 | 3.66% | 7402 | 3.64% | 453 | 3.58% | |
| 14 | 7335 | 3.32% | 7248 | 3.31% | 6738 | 3.32% | 456 | 3.61% | |
| 15 | 6595 | 2.99% | 6519 | 2.98% | 6082 | 2.99% | 423 | 3.34% | |
| 16 | 5603 | 2.54% | 5550 | 2.53% | 5091 | 2.51% | 349 | 2.76% | |
| 17 | 4285 | 1.94% | 4258 | 1.94% | 3952 | 1.95% | 348 | 2.75% | |
| 18 | 1690 | 0.77% | 1682 | 0.77% | 1569 | 0.77% | 117 | 0.92% | |
| NA | 11899 | 5.39% | 11755 | 5.37% | 10785 | 5.31% | 163 | 1.29% | |
| Asthma Intensity | | | | | | | | | |
| 1 | 155968 | 70.63% | 154648 | 70.61% | 143984 | 70.88% | 8959 | 70.83% | |
| 2 | 46322 | 20.98% | 45989 | 21.00% | 42377 | 20.86% | 2432 | 19.23% | |
| 3 | 18520 | 8.39% | 18373 | 8.39% | 16866 | 8.30% | 1258 | 9.95% | |

Sex
0 - "Indeterminate"
1 - "M", "Male"
2 - "F", "Female"
NA

Financial Class
1 - "MCD KS FAMILY HEALTH PARTNERS", "MCD KS FEE FOR SVC", "MCD KS MANAGED CARE",
2 - "MCD MO BLUE ADVANTAGE PLUS", "MCD MO FAMILY HEALTH PARTNERS", "MCD MO FEE FOR SERVICE", "MCD MO F
2 - "BLUE CROSS INDEMNITY", "BLUE CROSS MANAGED CARE", "CHAMPUS", "COMMERCIAL INS INDEMNITY", "COMMERC
4 - "SELF PAY"
5 - "UNKNOWN", "", "OTH CLIENT REFERRED", "RESEARCH/GRANT", "RX RETAIL PHARMACY"
NA

Race
1 - "CAUCASIAN/WHITE", "white"
2 - "Hispanic" (as race or ethnicity)
3 - "Black or African American", "BLACK"
4 - "American Indian or Alaska Native","Asian", "Multiracial","Native Hawaiian or Pacific Islander", "Other"
5 - " ", "Declined/Refused", "Respondent Not Available", "Unknown to Respondent"
NA

ICD9
1 - 493.00, 493.1, 493.90, 493.81, 493.82, 493.2
2 - 493.02, 493.12, 493.92, 493.22
3 - 493.01, 493.11, 493.91, 493.21