Review

Erika Arseneau and Cynthia M. Balion*

Statistical methods used in the calculation of geriatric reference intervals: a systematic review

DOI 10.1515/cclm-2015-0420
Received May 4, 2015; accepted July 27, 2015; previously published online August 26, 2015

Abstract

Background: Geriatric reference intervals (RIs) are not commonly available and are rarely used. It is difficult to select a reference population from a cohort with a high degree of morbidity. Also important are the statistical approaches used to determine health-associated reference values. It is the aim of this study to examine the statistical methods used in the calculation of geriatric RIs.

Methods: A search was conducted on EMBASE and Medline for articles between January 1989 and January 2014. Studies were selected if they: 1) were English primary articles; 2) performed a clinical chemistry test on a blood fraction; 3) had a population sub-group consisting of individuals ≥65 years of age; and 4) calculated a RI for the subgroup ≥65 years of age.

Results: There were 64 articles identified, of which 78.1% described the RI calculation method used. RI calculation was performed by non-parametric (21.9%), parametric (42.2%), robust (3.1%), or other (17.2%) methods. Outlier detection (SD, Grubb's test, Tukey's fence, Dixon) was infrequently used and although most studies performed partitioning, only 57.8% tested the statistical significance of the partitions. Few studies (17.2%) reported confidence intervals for the RI estimates. Overall, only 14.1% of studies provided RI estimates which followed the CLSI guideline EP28-A3c.

Conclusions: Statistical methods for RI calculation and partitioning varied considerably between studies and many failed to provide adequate descriptions of these methods. Challenges in analyses arose from insufficient sample sizes and heterogeneity in the elderly population. Geriatric RIs, although present in the literature, may not be properly calculated and should be carefully considered before applying them for clinical care.

Keywords: geriatric; reference interval; statistics; systematic review.

Introduction

Geriatric reference intervals (RIs) for laboratory tests are not commonly used. To aid in clinical decision-making, all laboratory tests require RIs to provide healthcare professionals with a normal range of values for comparison of patient test results [1]. Applicability of a RI is, however, dependent on how it was determined which includes factors, such as selection of the reference sample, sample size, analytical factors, such as instrumentation, biological factors, such as sex, age, demographic and lifestyle factors and even statistical approaches used in RI calculation [2].

Many methods for estimating RIs exist, however, there are three approaches that are most commonly used and are described in the Clinical and Laboratory Standards Institute (CLSI) guideline EP28-A3c for establishing RIs [3]. If the analyte measurements follow a Gaussian distribution, the parametric method of RI calculation may be applied [2]. This method computes the limits of a RI as mean±2 SD (2.5th and 97.5th percentiles) and historically is the most commonly used [4]. Unfortunately, many studies inappropriately apply this method to data where the normality assumption is not met [4]. If skewed, it is possible to transform the data, usually using a log transformation. This is simple to do mathematically, but log data is not intuitively easy to interpret [2]. Alternatively the non-parametric method, which does not hold any assumptions about the underlying distribution, can be used. With this method, the sample measurements are rank ordered and

***Corresponding author: Cynthia M. Balion,** Department of Laboratory Medicine, St. Joseph's Healthcare Hamilton, 50 Charlton Avenue E, Hamilton, ON L8N 4A6, Canada, Phone: +1 905 522 1155 ext 33098, Fax: +1 905 521 6165, E-mail: balion@hhsc.ca; Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario, Canada; and Department of Pathology and Molecular Medicine, McMaster University, Hamilton, Ontario, Canada
Erika Arseneau: Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario, Canada

the corresponding 2.5th and 97.5th percentiles are set as the RI [4]. The guideline suggests the use of the non-parametric approach to estimate the 2.5th and 97.5th percentiles for samples of ≥120 healthy subjects. In practice, obtaining a minimum of 120 samples can be very labor intensive and expensive [2]. This is particularly true for special populations, such as the elderly where morbidity is prevalent. It was reported by Horn and Pesce in 2003 that to obtain a 70–80-year-old healthy reference sample, nine out of every 10 people would have to be excluded [4]. When it is impossible to obtain the minimum number of samples, the guideline suggests the robust approach as an alternative method [3]. The robust approach is more statistically complex, involving an iterative procedure to estimate the median and median absolute deviation of the observed data [5].

The CLSI also provides guidelines for dealing with outlier data, partitioning the population to account for certain factors, such as age and sex, and for reporting RIs. Both outlier detection and partitioning should be done prior to calculating RI limits. Two different tests proposed by Dixon and Tukey respectively have been approved by the CLSI as methods for identifying outliers with the suggestion that identified outliers be removed [6, 7]. Partitioning is recommended by the CLSI as a method for determining RIs for different subclasses of the population, i.e. different sex or age groups. Tests, such as the ones by Harris and Boyd or Lahti, are suggested as methods for evaluating the statistical significance of these partitions [8, 9]. In terms of reporting RIs, it is also proposed that confidence intervals be provided to assess the precision of RI estimates [3]. These guidelines were developed primarily to establish a method of RI determination for adults and are silent on how to address concerns specific to elderly populations.

Identification of outliers, for instance, can become quite a cumbersome task in elderly populations where there is an apparent increase in biological variability [10]. It is plausible that it can become harder to discern which data points are true outliers when the sample is subject to more variability. In addition, the physiology of aging has a large impact on blood tests of the elderly. Geriatric patients commonly have one or more morbidity, may be taking multiple medications and have various ranges of physical ability, all affecting their biological states [11]. To effectively account for these factors a larger number of partitions would be necessary than would be for adult populations, making the minimum sample requirement even harder to obtain.

Before addressing the issues of applying current RI guidelines to geriatric populations it is first important to know what geriatric RIs are available in the literature and second to know what statistical methods have been used to determine them. A systematic review of past and current literature was performed to summarize what geriatric RIs are available. This paper specifically aims at examining the statistical methodologies used in these papers.

Materials and methods

A literature search on EMBASE and Medline databases was performed to identify articles published between January 1989 and January 2014. A preliminary literature search revealed an increase in the number of articles reporting elderly RIs published in 1989 into the early 1990s, validating the search start date. This preliminary literature search also identified useful search terms. Comprehensive searches were then performed using the terms ‘reference intervals’, ‘reference ranges’, ‘reference values’ and ‘reference parameters’ crossed with (operator AND) ‘elderly’, ‘old’, ‘geriatric’, ‘older adult’ with the field limit ‘humans’ (full search criteria located in Supplemental Table 1). A total of 985 articles, 982 with removal of duplicates, were found using these search criteria and imported into DistillerSR (Evidence Partners Incorporated, Ottawa, ON, Canada) for review.

Title and abstract screening was performed to select for articles that were in English, were primary research articles, performed a test on a blood fraction, measured a clinical chemistry analyte, and included people ≥65 years of age. Full text screening ensured the remaining articles had calculated a RI for at least one subgroup consisting only of individuals ≥65 years of age. The purpose of each study was identified using searches for the key terms ‘objective’, ‘aim’, ‘goal’ and/or ‘purpose’. Study purposes were then categorized into one of three groups: to establish RIs in general, to establish elderly RIs specifically, or to test a new analytical method of measurement for a given analyte.

Finally, data extraction was performed to extract all geriatric RIs, reference sample selection data, and information regarding analytical procedures. All data regarding statistical methods including the use of outlier detection, partitioning and partitioning tests, RI calculation and confidence interval reporting were also captured and are the focus of this article.

Results

The search strategy identified 985 articles. Title and abstract screening resulted in the selection of 344 studies (Figure 1). After full text screening a total of 100 studies (Figure 1) were

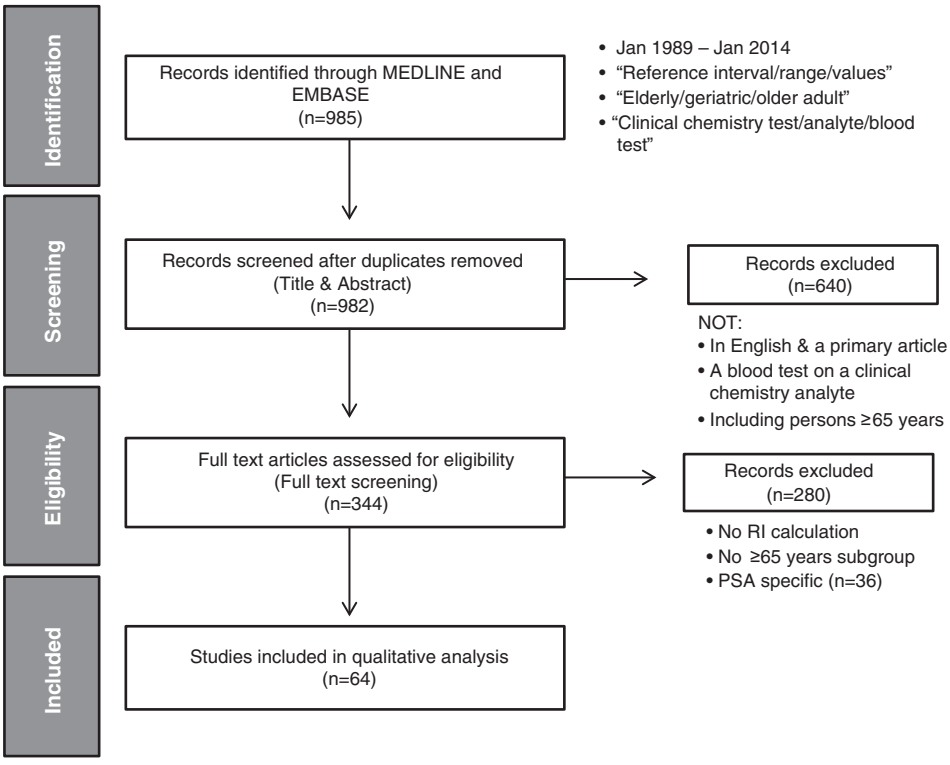

Figure 1: Analytical framework of the systematic review process.

found to have met the full text screening eligibility criteria. Of these studies there were 36 that described age-specific RIs for prostate-specific antigen (PSA). For the purposes of this systematic review these studies were excluded. Therefore a total of 64 papers were selected for final inclusion in the review (see Supplemental Table 2).

These studies were published in a variety of countries with the majority of publications (74.2%) coming from Europe and 21.0% coming from the US. Most studies were conducted to either establish a new method of analytical measurement (26.6%) or to calculate RIs in general (40.6%). Only one third of studies (35.9%) aimed to establish RIs specifically for the elderly and 29.7% looked only at people ≥65 years of age. In total 1094 geriatric RIs were captured from the 64 included studies. RIs were found for 94 analytes representing a broad range of physiological tests including markers of kidney and liver function, hormones, metabolites, lipids and enzymes.

A summary of the statistical methods used in establishing RIs are provided in Table 1. Detailed descriptions of the statistical methods for papers whose main purpose was to establish RIs specifically for the elderly (n=23) are outlined in Table 2. Detailed descriptions for all included articles can be found in Supplemental Table 2.

Only 34.4% of the 64 studies performed any type of outlier removal. The most common methods used were the standard deviation method (outside 3SD) (9.4%), the Grubbs' test (4.7%), Tukey's fence (6.3%), and Dixon criteria (6.3%). Some studies removed outliers based on more subjective criteria, e.g. Erdogan et al. [35] removed the top 10% of values for methylmalonic acid based on the fact that clinical histories associated with the samples obtained were unknown and there may be suspect disease within this tail-end population.

Most studies (92.2%) considered the need to look at the homogeneity of the data to determine if partitioning was needed (Table 1). Tests of significant differences between these partitions (primarily for sex and/or age) however, were conducted by only 57.8% of the studies. Despite using statistical tests of difference, only 10.9% collapsed insignificant partitions, 6.3% collapsed some insignificant partitions but not others, and 17.2% did not collapse insignificant partitions at all. For example, Shi et al. [36] looked at NT-proBNP and after partitioning age by decades found the only significant difference to be between persons aged 61–70 and aged 71–85 using the Kruskal-Wallis test, but in the end reported all RIs by decade. In contrast, another study that also estimated RIs for NT-proBNP, collapsed partitions that were proven to be significantly different. Alehagen et al. [12] after partitioning by 5-year age intervals demonstrated using an ANOVA with a post-hoc test, that NT-proBNP concentrations were significantly different

Table 1: Statistical methods used for reference interval determination in the 64 studies.

	% (n)[a]
Performed outlier detection	34.4 (22)
SD method (Outside 3SD)	9.4 (6)
Grubbs' test	4.7 (3)
Tukey's fence	6.3 (4)
Dixon	6.3 (4)
Subjective elimination[b]	7.8 (5)
Performed partitioning	92.2 (59)
Statistically tested partitioning	57.8 (37)
Collapsed insignificant partitions	
Yes	10.9 (7)
No	17.2 (11)
Some of the time	6.3 (4)
Described reference interval calculation methods[c]	78.1 (50)
Non-parametric	21.9 (14)
Appropriately (n≥120 per partition)	9.4 (6)
Inappropriately	12.5 (8)
Reported confidence intervals	9.4 (6)
Parametric	42.2 (27)
Applied to normally distributed/transformed data	23.4 (15)
Applied to skewed data	4.7 (3)
Underlying distribution unknown	14.1 (9)
Reported confidence intervals	6.25 (4)
Robust	3.1 (2)
Reported confidence intervals	3.1 (2)
Other[d]	17.2 (11)
Reported confidence intervals	1.6 (1)
Not reported	20.3 (13)
Reported confidence intervals	1.6 (1)
Reported confidence intervals	17.2 (11)
Referenced a RI guideline (IFCC, CLSI/NCCLS)	40.6 (26)
Followed the CLSI guideline[e]	12.5 (8)

[a]Some studies used more than one statistical approach and may therefore be counted in more than one method category; [b]Refers to removal of outliers by examination of plots or tail end data (usually with clinical basis for removal); [c]Includes studies that specifically stated method type and those that implied the method or were able to be deduced by references; [d]Used reference curves or did not report enough detail to classify the method type; [e]Defined as properly using one or more methods for establishing a RI as described by the CLSI and reporting confidence intervals for the associated estimates. CLSI, Clinical Laboratory Standards Institute (formerly known as NCCLS (National Committee for Clinical Laboratory Standards); IFCC, International Federation of Clinical Chemistry.

in persons aged 71–75 (n=120) compared with those older than 75 years (n=53). However, only one RI for NT-proBNP was reported (≥65 years) due to sample size restraints of the 5-year age intervals [12]. Overall, there was inconsistency in how partitions were dealt with in regards to the number of samples per partition and their significance.

Methods for the calculation of RIs substantially varied between studies and some studies employed more than one method (Supplemental Table 2). Fourteen studies [14, 19, 21, 22, 30, 32, 33, 35, 37–42] implemented the rank-based non-parametric approach recommended by the CLSI although only six [14, 30, 32, 38, 40, 42] of these studies abided by the minimum requirement of 120 subjects per partition. Six [14, 22, 30, 32, 39, 40] of the 14 studies also provided confidence intervals for their lower and upper limit estimates.

Twenty-seven studies [18, 20, 22–26, 29, 31, 41, 43–59] used the parametric method for RI determination. Fifteen studies [23, 25, 29, 31, 43–48, 50, 53–54, 56, 59] applied the method appropriately to normally distributed values or values that were transformed to approximate a normal distribution. Three studies [18, 41, 49] applied a parametric calculation to skewed data. The remaining nine studies [20, 22, 24, 26, 51, 52, 55, 57, 58] did not specifically state the underlying distribution of the data making it impossible to tell whether the parametric method was applied appropriately or not. In 15 studies [22–25, 31, 41, 44–46, 49, 50, 53–56] calculations were made appropriately using the mean±2 SD (or 2.5th and 97.5th percentiles). Other methods of RI determination included using the mean±1.65 SD [20], mean±SE [47, 52], mean±SD [26, 29, 43, 48, 51, 58], 5th and 95th percentile [18, 25, 48, 52, 57, 58] or 10th and 90th percentile [51]. Several studies reported RIs using more than one method [25, 46, 48, 51, 52, 58]. Only four [22, 25, 57, 59] of the 26 studies provided confidence intervals for their estimates.

Two studies [14, 30] employed the robust method and both provided confidence intervals. Eight studies [60–67] employed reference curves to estimate RIs and only one study [60] calculated confidence intervals. Three studies [68–70] used other methods to determine RIs. The remaining 13 studies [12, 13, 15–17, 27, 34, 36, 71–75] did not provide enough specific details to classify the type of RI calculation method used and only one study [16] provided confidence intervals for their estimates.

Less than half of the studies included in this review (40.6%) referenced any guideline for calculating RIs with very few articles (12.5%) directly citing a RI guideline and ensuring the appropriate use of its methods. Twenty-six articles of the included articles were published in or after 2008, when the most recent CLSI guideline was published, however, only three of those used statistical methods that abide by the CLSI criteria. More importantly, only eight studies (12.5%) from all that were included were found to have reported statistically valid RIs that follow current CLSI criteria [14, 22, 25, 30, 32, 40, 57, 59]. Studies were classified as satisfying CLSI criteria if they explicitly stated the type of method used, applied parametric or non-parametric

Table 2: Statistical methods and results of the 23 studies whose main purpose was to calculate geriatric reference intervals.

Author Year	Outlier detection method	Partitioning method[a]	Partition types[b]			Partitions ≥120, %	Data distribution(s)[c]	Reference interval method[d]	Reference interval type[e]	Confidence interval	Guideline cited	Methods abide CLSI criteria[f]
			Age, years	Sex	Other							
Alehagen 2007 [12]	Horn's algorithm with 1.5 fence factor	ANOVA and Tukey post-hoc for age; significant but did not partition	>65			0	Skewed, log transformed	NR, bootstrap estimation	2.5th and 97.5th percentiles	No	IFCC	No
Boulat 1998 [13]	Tukey's fence; Cross-referenced with clinical data	Mann-Whitney U-test for sex; significant for some analytes, collapsed insignificant	>65	Y		75	Skewed distribution for most analytes	NR	2.5th and 97.5th percentiles	No	IFCC	No
Carlsson 2010 [14]	NR	Mann-Whitney U significant for some analytes, did not collapse insignificant	70	Y	CVD	100	NR	*Non-parametric and robust, Bootstrap estimation	2.5th and 97.5th percentiles	Yes, 90%	IFCC	Yes
Erasmus 1997 [15]	3 SD method	Unpaired t-test for sex; significant	10 year intervals	Y		0	NR	NR	NR	No	IFCC	No
Eskelinen 2005 [16]	3 SD method	Mann-Whitney U-test for sex; significant	>65	Y	TgAb, TPOAb	100	Skewed, log transformed	NR	2.5th and 97.5th percentiles	Yes, 95%	No	No
Evrin 2005 [17]	NR	t-test for sex; did not specify result	>70			NR	Skewed, log transformed	NR	NR	No	No	No
Garry 1989 [18]	NR	NR	>65			NR	Normal, for most but some skewed	Parametric, normal distribution approximation	5th and 95th percentiles	No	No	No
Hammerman-Rozenberg 1996 [19]	Dixon; w/Reed criterion	NR	10 year intervals	Y		50	NR	Non-parametric, percentile estimation method	2.5th and 97.5th percentiles	No	No	No
Hardie 2004 [20]	Subjective elimination; graphically using scatter plot	t-test for sex; significant	>70	Y		0	NR	Parametric	Mean±1.65 SD	No	No	No
Herbeth 2001 [21]	NR	NR; significant for age not sex, collapsed insignificant	50–64, 65–99	Y		0	NR, log transformed	Non-parametric	95th percentile	No	IFCC	No

Table 2 (continued)

Author Year	Outlier detection method	Partitioning method[a]	Partition types[b]			Partitions ≥120, %	Data distribution(s)[c]	Reference interval method[d]	Reference interval type[e]	Confidence interval	Guideline cited	Methods abide CLSI criteria[f]
			Age, years	Sex	Other							
Huber 2006 [22]	Subjective elimination; Visual assessment and clinical evidence	Lahti for sex; significant for some, collapsed insignificant	75	Y		34.7	Log-normal, most tests	Non-parametric when n≥120, parametric when n<120	2.5th and 97.5th percentiles	Yes, 90%	IFCC	Yes
Joosten 1996 [23]	NR	NR	>65	Y		0	NR, log transformed	Parametric	Mean±2 SD	No	No	No
Kubota 2012 [24]	NR	t-test for age and sex; significant for both	10 year intervals, and >85			100	NR	Parametric	Mean±2 SD	No	NCCLS	No
Lawrence 1991 [25]	NR	NR	>65	Y		NR	Normal (some); Skewed (some), log transformed	Parametric	2.5th and 97.5th percentiles and 5th and 95th percentiles	Yes, SD of associated percentiles	No	Yes
Lio 2008 [26]	NR	t-test for age; some significant, did not collapse insignificant	65–85, and >100	Y		100	NR	Parametric	Mean±2 SD	No	No	No
Maugeri 1996 [27]	NR	NR	>65			100	NR	NR	NR	No	No	No
Millan-Calenti 2012 [28]	Subjective elimination; More than 5% above upper limit or below lower limit	NR	>65	Y		100	NR	NR	2.5th and 97.5th percentiles	No	No	No
Robbins 1995 [29]	NR	Two-factor ANOVA for age and sex; significant	10 year intervals, and >75	Y		100	Normal (some); Skewed (some), log transformed	Parametric	Mean±SD	No	IFCC	No
Ryden 2012 [30]	NR	Mann-Whitney U for presence/absence CVD; significant for some analytes, did not collapse insignificant	70	Y	CVD	100	NR	*Non-parametric and robust, bootstrap estimation	2.5th and 97.5th percentiles	Yes (90%)	IFCC	Yes

Table 2 (continued)

Author Year	Outlier detection method	Partitioning method[a]	Partition types[b]			Partitions ≥120, %	Data distribution(s)[c]	Reference interval method[d]	Reference interval type[e]	Confidence interval	Guideline cited	Methods abide CLSI criteria[f]
			Age, years	Sex	Other							
Stulnig 1993 [31]	NR	Mann-Whitney U-test for age and sex; some significant, collapsed insignificant sex differences but not age differences	>65	Y		0	Skewed (some), log transformed	Parametric	Mean±1.96 SD	No	IFCC	No
Takala 2002 [32]	Dixon	Wilcoxon rank-sum for sex; significant	>65	Y	Iron status	100	NR	Non-parametric, bootstrap estimation	2.5th and 97.5th percentiles	Yes, 90%	IFCC	Yes
Tietz 1992 [33]	Dixon; w/Reed criterion	t-test for age and sex; some significant, collapsed insignificant sex differences but not age differences	60–90, and >90	Y		13.5	NR	Non-parametric	2.5th and 97.5th percentiles	No	IFCC	No
Yeap 2012 [34]	NR	NR	>70	M		100	Skewed (some)	NR	2.5th and 97.5th percentiles	No	No	No

CLSI, Clinical Laboratory Standards Institute; CVD, cardiovascular disease; IFCC, International Federation of Clinical Chemistry and Laboratory Medicine; NR, not reported; RI, reference interval; TgAb, thyroglobulin antibody; TPOAb, thyroid peroxidase antibodies. [a]The results of the partitioning, if available, are given. The Mann-Whitney U-test is comparable to the Wilcoxon Rank Sum Test; [b]Partitions listed are those reported and may not be consistent with the results of the partitioning tests; [c]Reflects distributions of all analytes reported in the study; [d]The non-parametric method refers to the rank-based procedure described by CLSI. The parametric method refers to reporting RIs as mean±2 SD. The robust method refers to the iterative bootstrapping method also described by the CLSI. Any other method used was classified as other. An asterisk (*) indicates the method type was not specifically stated within the paper but was able to be deduced or assumed by the information provided; [e]Mean±1.65 SD is equivalent to 5th and 95th percentiles and mean±2 SD is equivalent to 2.5th and 97.5th percentiles; [f]Defined as properly using one or more methods for establishing a RI as described by the CLSI and reporting confidence intervals for the associated estimates.

methods appropriately, and provided confidence intervals for their estimates.

A prime example of application of the CLSI guideline was the study by Huber et al. [22]. In addition to following the CLSI guideline as defined above this study also removed outliers and tested for significant differences between sex partitions using the Lahti method, collapsing those that were not significant. They also demonstrated a different approach to calculating geriatric RIs, attempting to eliminate the effect of aging by examining only 75 year olds [22].

Discussion

This systematic review provided a comprehensive look at statistical methods that have been used to calculate geriatric RIs over the last 25 years. Evaluation of the 64 studies included in this review revealed gaps in reporting statistical methods used for calculating RIs and highlights the difficulties in applying current RI guidelines to geriatric populations. Attention to the high prevalence of disease and heterogeneity in biological aging in an older population, and limited sample sizes, was not addressed well in the majority of studies.

There are many steps in the process of determining RIs.The first common practice step to RI determination is the removal of outliers even when working with a healthy population. This helps to eliminate subjects with underlying undiagnosed conditions. However, it is not quantified how much certain analytes change as part of the 'normal' aging process. Certain trends toward increasing and decreasing values have been identified, e.g. the increase of creatinine with age [44], but no limits have been established for what would be considered a normal increase vs. an increase that is indicative of disease. Therefore when examining outlier removal in the elderly it is important to remember that outliers may represent a natural variability within a given group of individuals [5]. A method to assess variability, such as sensitivity analysis could be performed to determine how influential outlying observations are on RI estimates [5].

Furthermore, exclusion of outliers when using hospital or laboratory databases is a special case and statistical approaches, such as the Bhattacharyya method should be used to select 'healthy' individuals from these data sets to assure more reliable and valid RIs. Six studies [35, 38, 40, 50, 71, 75] used these datasets, but only one study [40] used the Bhattacharyya method. The selection of reference participants from these databases is not ideal or recommended by the CLSI, but may be useful for estimating RIs of hard to reach populations, such as the elderly.

In a recent systematic review examining statistical methods used for pediatric RIs [5] it was shown that the two most common outlier detection methods were the Dixon method and the Tukey method. The Dixon method compares D (the absolute difference between extreme observations and the next largest/smallest observation) and R (the range of all observations) to determine whether outliers exist in the data set [3]. Commonly, the Reed criteria is applied which suggests a ≥1/3 cut-off for D/R. This limit is considered conservative for large samples but may fail when >2–3 outliers are present [3] which would not be uncommon in a geriatric population. The Tukey method consists of excluding outliers if they fall outside of 1.5 times the interquartile range [3]. Both methods are based under the normality assumption [5] and given that most analyte data for the elderly is skewed [11] a different method of outlier detection may be necessary.

Following outlier detection, data is partitioned into homogenous groups to reflect biological variability properly. For age, this is usually done by predefining chronological age categories for partitioning using 5- or 10-year age intervals. Using this method for elderly populations is difficult as it is hard to obtain large numbers of healthy elderly persons, especially when defining 'healthy' as the absence of disease which is typically done for RI studies. Furthermore, this categorical age partitioning may not be suitable for elderly populations given that one's chronological age may not be indicative of their biological state. Applying standard age partitioning to elderly persons results in grouping a number of people with heterogeneous health states together. For instance, consider even selecting two 70-year-old males in relatively good health. One is fairly mobile and walks without assistance compared to the other who is dependent on a wheelchair for day-to-day movement. This simple difference in mobility may alter the biological status of various analytical markers, yet standard partitioning methods would classify these individuals into the same reference population.

Rather than continuing the common practice of using categorical age partitions for the elderly it may be more useful to compare RIs for groups of individuals that are of similar biological states. To do this, it may be more worthwhile to consider visually assessing the data for more homogenous groups. This visual examination would allow researchers to identify groups that have similar laboratory values and can be done by using simple scatter or box plots [5] against age, gender, number and/or type of

morbidity or any other covariates to identify changes in any given analyte.

Currently, there is no guideline that provides advice on how to choose the appropriate method to test significant differences between partitions, though it is mentioned by the CLSI guideline as a step that should be considered in RI determination. Common statistical tests used for this purpose and comparison of means and/or medians include the t- and F- tests for Gaussian data or the Mann-Whitney U-Test (or Wilcoxon Rank Sum Test) for skewed data. Alternatively, calculating all potential partitions and testing the resulting limits of different groups by the method of Lahti et al. [9] or Sinton et al. [76] could be done. Regardless of the method, insignificant partitions should be collapsed and significant partitions should be kept separate. Unfortunately, as described by Shi et al. [36] and Alehagen et al. [12], it is difficult to do either with elderly populations. Both cases demonstrated that NT-proBNP concentrations change with age and both exemplify the difficulties that arise due to difficulties in obtaining sufficient sample sizes for age groups ≥65 years of age. In one case you have partitions appropriately set based on statistical tests but the sample sizes are inadequate. In the other you have two very different groups being grouped together to attain the recommended sample sizes. Neither case adequately portrays the differences in age with proper methods.

The final step in RI determination is the estimation of the lower and upper limits. When selecting which statistical method to use for this estimation it is important to consider the underlying assumptions and recommendations for each method. This review found that methods for calculating RIs significantly differed between studies but more importantly that the underlying distribution of the data often went unreported or was not considered in the approach that was chosen. Simple visualization of a frequency plot or a normal probability plot allows for assessment of the distribution and skewness of the data before deciding on a calculation method [5].

Sample size was also not often considered when choosing the RI calculation method. This was evidenced by the number of articles that used the non-parametric method for partitions with fewer than 120 samples and little use of the robust approach for small sample sizes. The reason for limited use of the robust method could be due to its statistical complexity, although statistical software now exists that can accommodate these types of analyses.

The paucity of studies reporting confidence intervals for RI limits is a problem. This is because the width of a confidence interval indicates how precise RI estimates are and provides awareness of sampling variability. It is an important parameter for a geriatric population given the heterogeneity of this population even in the 'healthy' group. The precision of confidence intervals can be improved by increasing sample sizes, provided the sample is homogenous. Indeed it may need to be much greater than the recommended 120 samples to increase the precision of the calculated limits. The absence of confidence intervals also restricts any meta-analyses that could be performed across different studies for the same analyte. Furthermore, little work has been done to evaluate the impact of each calculation method on RI estimates [5]. Simulation studies to investigate this and the effects of outlier detection and partitioning methods are needed.

This systematic review was limited in scope to only published primary research studies although it is recognized that RI studies are often performed as part of requirements for clinical laboratory accreditation and often go unpublished. However, it is unlikely that very many have looked specifically at the geriatric population given that geriatric RIs are rarely used clinically. Furthermore, evaluation of study quality was not performed except in the limited sense of determining if they used criteria outlined in the RI guideline from CLSI [3]. There are currently no quality assessment tools available to evaluate the quality of RI studies.

In summary, there are relatively few published studies specific to geriatric RIs as compared to the adult population and the statistical approaches of their calculations are varied. This is in part due to the absence of appropriate statistical methods and guidance specific to this heterogeneous population. Descriptions of methods used in RI studies are also problematic in that incomplete information is provided, making it difficult to understand exactly how analyses are performed. Validity of geriatric RIs however, is not only based on appropriate statistical methodology but also on appropriateness of participant selection. This systematic review, focused only on the statistical methodology, found that most studies have failed to analyze the data correctly when estimating geriatric RIs. This highlights the need for improvement in the field of RI methodology, particularly for this unique population.

Acknowledgments: The authors would like to thank Maureen Rice and Mary Gauld for aiding in the library search. Thank you to Dr. Jemila Hamid, Dr. Parminder Raina and Caitlin Daly for their guidance and support. This research was funded by the CLSA Mobility Initiative (CMI) (MT 92026) – An Emerging Team in Mobility grant, and graduate and research scholarships from McMaster

University and the Department of Clinical Epidemiology & Biostatistics.

Author contributions: All the authors have accepted responsibility for the entire content of this submitted manuscript and approved submission.

Research funding: None declared.

Employment or leadership: None declared.

Honorarium: None declared.

Competing interests: The funding organization(s) played no role in the study design; in the collection, analysis, and interpretation of data; in the writing of the report; or in the decision to submit the report for publication.

References

1. Faulkner WR, Demers LM. Importance of age-dependent reference values in biochemical testing: are we including the elderly population? Clin Chem 1994;40:855–6.
2. Horn PS, Pesce AJ. Reference intervals: a user's guide. Washington, DC: AACC Press, 2005.
3. CLSI. Defining, establishing and verifying reference intervals in the clinical laboratory; Approved Guideline, 3rd ed. CLSI document EP28-A3c. Wayne, PA: Clinical and Laboratory Standards Institute, 2008.
4. Horn PS, Pesce AJ. Reference intervals: An update. Clin Chim Acta 2003;334:5–23.
5. Daly CH, Liu X, Grey VL, Hamid JS. A systematic review of statistical methods used in constructing pediatric reference intervals. Clin Biochem 2013;46:1220–7.
6. Dixon WJ. Processing data for outliers. Biometrics 1953;9:74–89.
7. Tukey JW. Exploratory data analysis. Reading: Addison-Wesley Publishing Company 1977;2:688–695.
8. Harris EK, Boyd JC. On dividing reference data into subgroups to produce separate reference ranges. Clin Chem 1990;36:265–70.
9. Lahti A, Petersen PH, Boyd JC, Rustad P, Laake P, Solberg HE. Partitioning of nongaussian-distributed biochemical reference data into subgroups. Clin Chem 2004;50:891–900.
10. Kruger A. The limits of normality in elderly patients. Bailliere's Clinical Haematology 1987;1:271–89.
11. Faulkner W, Meites S, editors. Geriatric clinical chemistry: reference values. Washington DC: AACC Press, 1994.
12. Alehagen U, Goetze JP, Dahlström U. Reference intervals and decision limits for B-type natriuretic peptide (BNP) and its precursor (Nt-proBNP) in the elderly. Clin Chim Acta 2007;382:8–14.
13. Boulat O, Krieg MA, Janin B, Burckhardt P, Francioli P, Bachmann C. Clinical chemistry variables in normal elderly and healthy ambulatory populations: comparison with reference values. Clin Chim Acta 1998;272:127–35.
14. Carlsson L, Lind L, Larsson A. Reference values for 27 clinical chemistry tests in 70-year-old males and females. Gerontology 2010;56:259–65.
15. Erasmus RT, Ray U, Nathaniel K, Dowse G. Reference ranges for serum creatinine and urea in elderly coastal Melanesians. P N G Med J 1997;40:89–91.
16. Eskelinen S, Suominen P, Vahlberg T, Löppönen M, Isoaho R, Kivelä SL, et al. The effect of thyroid antibody positivity on reference intervals for thyroid stimulating hormone (TSH) and free thyroxine (FT4) in an aged population. Clin Chem Lab Med 2005;43:1380–5.
17. Evrin PE, Nilsson SE, Oberg T, Malmberg B. Serum C-reactive protein in elderly men and women: association with mortality, morbidity and various biochemical values. Scand J Clin Lab Invest 2005;65:23–31.
18. Garry PJ, Hunt WC, VanderJagt DJ, Rhyne RL. Clinical chemistry reference intervals for healthy elderly subjects. Am J Clin Nutr 1989;50:1219–30.
19. Hammerman-Rozenberg R, Cohen A, Ginsberg G, Maaravi Y, Ebstein RP, Stessman J. Laboratory reference values for the 70 year olds. Isr J Med Sci 1996;32:611–20.
20. Hardie JA, Vollmer WM, Buist S, Ellingsen I, Morkve O. Reference values for arterial blood gases in elderly. CHEST J 2004;125:2053–60.
21. Herbeth B, Siest G, Henny J. High sensitivity C-reactive protein (CRP) reference intervals in the elderly. Clin Chem Lab Med 2001;39:1169–70.
22. Huber KR, Mostafaie N, Stangl G, Worofka B, Kittl E, Hofmann J, et al. Clinical chemistry reference values for 75-year-old apparently healthy persons. Clin Chem Lab Med 2006;44:1355–60.
23. Joosten E, Lesaffre E, Riezler R. Are different reference intervals for methylmalonic acid and total homocysteine necessary in elderly people? Eur J Haematol 1996;57:222–6.
24. Kubota K, Kadomura T, Ohta K, Koyama K, Okuda H, Kobayashi M, et al. Analyses of laboratory data and establishment of reference values and intervals for healthy elderly people. J Nutr Heal Aging 2012;16:412–6.
25. Lawrence CJ, Trewin VF. The construction of biochemical reference ranges and the identification of possible adverse drug reactions in the elderly. Stat Med 1991;10:831–7.
26. Lio D, Malaguarnera M, Maugeri D, Ferlito L, Bennati E, Scola L, et al. Laboratory parameters in centenarians of Italian ancestry. Exp Gerontol 2008;43:119–22.
27. Maugeri D, Russo MS, Carnazzo G, Di Stefano F, Catanzaro S, Campagna S, et al. Altered laboratory thyroid parameters indicating hyperthyroidism in elderly subjects. Arch Gerontol Geriatr 1996;22:145–53.
28. Millán-Calenti JC, Sánchez A, Lorenzo-López L, Maseda A. Laboratory values in a Spanish population of older adults: a comparison with reference values from younger adults. Maturitas 2012;71:396–401.
29. Robbins J, Wahl P, Savage P, Enright P, Powe N, Lyles M. Hematological and biochemical laboratory values in older Cardiovascular Health Study participants. J Am Geriatr Soc 1995;43:855–9.
30. Ryden I, Lind L, Larsson A. Reference values of thirty-one frequently used laboratory markers for 75-year-old males and females. Ups J Med Sci 2012;117:264–72.
31. Stulnig T, Mair A, Jarosch E, Schober M, Schönitzer D, Wick G, et al. Estimation of reference intervals from a SENIEUR protocol compatible aged population for immunogerontological studies. Mech Ageing Dev 1993;68:105–15.
32. Takala TI, Suominen P, Isoaho R, Kivela S-L, Lopponen M, Peltola O, et al. Iron-replete reference intervals to increase sensistivity of hematologic and iron status laboratory tests in the elderly. Clin Chem 2002;48:1586–9.
33. Tietz NW, Shuey DF, Wekstein DR. Laboratory values in fit aging individuals – sexagenarians through centenarians. Clin Chem 1992;38:1167–85.

34. Yeap BB, Alfonso H, Chubb SA, Handelsman DJ, Hankey GJ, Norman PE, et al. Reference ranges and determinants of testosterone, dihydrotestosterone, and estradiol levels measured using liquid chromatography-tandem mass spectrometry in a population-based cohort of older men. J Clin Endocrinol Metab 2012;97:4030–9.
35. Erdogan E, Nelson GJ, Rockwood AL, Frank EL. Evaluation of reference intervals for methylmalonic acid in plasma/serum and urine. Clin Chim Acta 2010;411:1827–9.
36. Shi X, Xu G, Xia T, Song Y, Lin Q. N-terminal-pro-B-type natriuretic peptide (NT-proBNP): reference range for Chinese apparently healthy people and clinical performance in Chinese elderly patients with heart failure. Clin Chim Acta 2005;360:122–7.
37. Bissé E, Epting T, Beil A, Lindinger G, Lang H, Wieland H. Reference values for serum silicon in adults. Anal Biochem 2005;337:130–5.
38. Bock BJ, Dolan CT, Miller GC, Fitter WF, Hartsell BD, Crowson AN, et al. The data warehouse as a foundation for population-based reference intervals. Am J Clin Pathol 2003;120:662–70.
39. Eskelinen S, Vahlberg T, Isoaho R, Kivelä SL, Irjala K. Biochemical reference intervals for sex hormones with a new AutoDelfia method in aged men. Clin Chem Lab Med 2007;45:249–53.
40. Mold JW, Aspy CB, Blick KE, Lawler FH. The determination and interpretation of reference intervals for multichannel serum chemistry tests. J Fam Pract 1998;46:223–41.
41. Pottel H, Vrydags N, Mahieu B, Vandewynckele E, Croes K, Martens F. Establishing age/sex related serum creatinine reference intervals from hospital laboratory data based on different statistical methods. Clin Chim Acta 2008;396:49–55.
42. Wahlin A, Backman L, Hultdin J, Adolfsson R, Nilsson L-G. Reference values for serum levels of vitamin B 12 and folic acid in a population-based sample of adults between 35 and 80 years of age. Public Health Nutr 2002;5:505–11.
43. Abiaka C, Olusi S, Al-Awadhi A. Reference ranges of copper and zinc and the prevalence of their deficiencies in an Arab population aged 15–80 years. Biol Trace Elem Res 2003;91:33–43.
44. Andreassen M, Nielsen K, Raymond I, Kristensen LØ, Faber J. Characteristics and reference ranges of Insulin-Like Growth Factor-I measured with a commercially available immunoassay in 724 healthy adult Caucasians. Scand J Clin Lab Invest 2009;69:880–5.
45. Duncanson GO, Worth HG. Determination of reference intervals for serum magnesium. Clin Chem 1990;36:756–8.
46. Elmlinger MW, Kühnel W, Weber MM, Ranke MB. Reference ranges for two automated chemiluminescent assays for serum insulin-like growth factor I (IGF-I) and IGF-binding protein 3 (IGFBP-3). Clin Chem Lab Med 2004;42:654–64.
47. Ganji V, Kafai MR. Population reference values for plasma total homocysteine concentrations in US adults after the fortification of cereals with folic acid. Am J Clin Nutr 2006;84:989–94.
48. Jungner I, Marcovina SM, Walldius G, Holme I, Kolar W, Steiner E. Apolipoprotein B and A-1 values in 147 576 Swedish males and females, standardized according to the World Health Organization – International Federation of Clinical Chemistry First International Reference Materials. Clin Chem 1998;44:1641–9.
49. Kahapola-Arachchige KM, Hadlow N, Wardrop R, Lim EM, Walsh JP. Age-specific TSH reference ranges have minimal impact on the diagnosis of thyroid dysfunction. Clin Endocrinol (Oxf) 2012;77:773–9.
50. Kairisto V, Hänninen KP, Leino A, Pulkki K, Peltola O, Näntö V, et al. Generation of reference values for cardiac enzymes from hospital admission laboratory data. Eur J Clin Chem Clin Biochem 1994;32:789–96.
51. Kornitzer M, Bara L. Clinical and anthropometric data, blood chemistry and nutritional patterns in the Belgian population according to age and sex. Acta Cardiol 1989;44:101–44.
52. Najjar M, Carter-Pokaras O. Clinical chemistry profile data for Hispanics 1982–84. National Center for Health Statistics. Vital Heal Stat 1992;11:1–53.
53. Ritchie RF, Palomaki GE, Neveux LM, Navolotskaia O, Ledue TB, Craig WY. Reference distributions for the positive acute phase serum proteins, alpha1-acid glycoprotein (orosomucoid), alpha1-antitrypsin, and haptoglobin: a practical, simple, and clinically relevant approach in a large cohort. J Clin Lab Anal 2000;14:284–92.
54. Ritchie RF, Palomaki GE, Neveux LM, Navolotskaia O, Ledue TB, Craig WY. Reference distributions for serum iron and transferrin saturation: a practical, simple, and clinically relevant approach in a large cohort. J Clin Lab Anal 2002;16:237–45.
55. Sokoll LJ, Russell RM, Sadowski JA, Morrow FD. Establishment of creatinine clearance reference values for older women. Clin Chem 1994;40:2276–81.
56. Vicente C, Porto G, De Sousa M. Method for establishing serum ferritin reference values depending on sex and age. J Lab Clin Med 1990;116:779–84.
57. Wener MH, Daum Phyllis R, McQuillan GM. The Influence of age, sex, and race on the upper reference limit of serum C-reactive protein concentration. J Rheumatol 2000;27:2351–9.
58. Wu TL, Chen TI, Chang PY, Tsao KC, Sun CF, Wu LL, et al. Establishment of an in-house ELISA and the reference range for serum amyloid A (SAA). Complementarity between SAA and C-reactive protein as markers of inflammation. Clin Chim Acta 2007;376:72–6.
59. Friis-Hansen L, Hilsted L. Reference intervals for thyreotropin and thyroid hormones for healthy adults based on the NOBIDA material and determined using a Modular E170. Clin Chem Lab Med 2008;46:1305–12.
60. Bjerner J, Høgetveit A, Wold Akselberg K, Vangsnes K, Paus E, Bjøro T, et al. Reference intervals for carcinoembryonic antigen (CEA), CA125, MUC1, Alfa-foeto-protein (AFP), neuron-specific enolase (NSE) and CA19.9 from the NORIP study. Scand J Clin Lab Invest 2008;68:703–13.
61. Bjerner J, Biernat D, Fosså SD, Bjøro T. Reference intervals for serum testosterone, SHBG, LH and FSH in males from the NORIP project. Scand J Clin Lab Invest 2009;69:873–9.e1–11.
62. Brabant G, Von Zur Mühlen A, Wüster C, Ranke MB, Kratzsch J, Kiess W, et al. Serum insulin-like growth factor I reference values for an automated chemiluminescence immunoassay system: results from a multicenter study. Horm Res Paediatr 2003;60:53–60.
63. Friedrich N, Völzke H, Rosskopf D, Steveling A, Krebs A, Nauck M, et al. Reference ranges for serum dehydroepiandrosterone sulfate and testosterone in adult men. J Androl 2008;29:610–7.
64. Friedrich N, Alte D, Völzke H, Spilcke-Liss E, Lüdemann J, Lerch MM, et al. Reference ranges of serum IGF-1 and IGFBP-3 levels in a general adult population: results of the Study of Health in Pomerania (SHIP). Growth Horm IGF Res 2008;18:228–37.

65. Friedrich N, Krebs A, Nauck M, Wallaschofski H. Age- and gender-specific reference ranges for serum insulin-like growth factor I (IGF-I) and IGF-binding protein-3 concentrations on the Immulite 2500: results of the Study of Health in Pomerania (SHIP). Clin Chem Lab Med 2010;48:115–20.
66. Haring R, Hannemann A, John U, Radke D, Nauck M, Wallaschofski H, et al. Age-specific reference ranges for serum testosterone and androstenedione concentrations in women measured by liquid chromatography-tandem mass spectrometry. J Clin Endocrinol Metab 2012;97:408–15.
67. Salameh WA, Redor-Goldman MM, Clarke NJ, Reitz RE, Caulfield MP. Validation of a total testosterone assay using high-turbulence liquid chromatography tandem mass spectrometry: total and free testosterone reference ranges. Steroids 2010;75:169–75.
68. Bayram F, Gedik VT, Demir Ö, Kaya A, Gündoan K, Emral R, et al. Epidemiologic survey: reference ranges of serum insulin-like growth factor 1 levels in Caucasian adult population with immunoradiometric assay. Endocrine 2011;40:304–9.
69. Contois JH, McNamara JR, Lammi-Keefe CJ, Wilson PW, Massov T, Schaefer EJ. Reference intervals for plasma apolipoprotein B determined with a standardized commercial immunoturbidimetric assay: results from the Framingham Offspring Study. Clin Chem 1996;42:515–23.
70. Eskes SA, Tomasoa NB, Endert E, Geskus RB, Fliers E, Wiersinga WM. Establishment of reference values for endocrine tests. Part VII: growth hormone deficiency. Neth J Med 2009;67:127–33.
71. Krøll J, Saxtrup O. On the use of patient data for the definition of reference intervals in clinical chemistry. Scand J Clin Lab Invest 1998;58:469–74.
72. Li C, Guan H, Teng X, Lai Y, Chen Y, Yu J, et al. An epidemiological study of the serum thyrotropin reference range and factors that influence serum thyrotropin levels in iodine sufficient areas of China. Endocr J 2011;58:995–1002.
73. Meinitzer A, Puchinger M, Winklhofer-Roob BM, Rock E, Ribalta J, Roob JM, et al. Reference values for plasma concentrations of asymmetrical dimethylarginine (ADMA) and other arginine metabolites in men after validation of a chromatographic method. Clin Chim Acta 2007;384:141–8.
74. Takeda K, Mishiba M, Sugiura H, Nakajima A, Kohama M, Hiramatsu S. Evaluated reference intervals for serum free thyroxine and thyrotropin using the conventional outliner rejection test without regard to presence of thyroid antibodies and prevalence of thyroid dysfunction in Japanese subjects. Endocr J 2009;56:1059–66.
75. Vadiveloo T, Donnan PT, Murphy MJ, Leese GP. Age- and gender-specific TSH reference intervals in people with no obvious thyroid disease in tayside, Scotland: the thyroid epidemiology, audit, and research study (TEARS). J Clin Endocrinol Metab 2013;98:1147–53.
76. Sinton TJ, Cowley DM, Bryant SJ. Reference intervals for calcium, phosphate, and alkaline phosphatase as derived on the basis of multichannel-analyzer profiles. Clin Chem 1986;32:76–9.

Supplemental Material: The online version of this article (DOI: 10.1515/cclm-2015-0420) offers supplementary material, available to authorized users.