Methods

1. Protocol and registration

This systematic review and meta-analysis was carried out according to the PRISMA-DTA guidelines (Preferred Reporting Items for Systematic Reviews and Meta-analysis of Diagnostic Test Accuracy Studies -The PRISMA-DTA Statement) (15) for the abstract and the body of the manuscript (S1 and S2). The protocol was registered in the PROSPERO database (International Prospective Register of Systematic Reviews) with number CRD42020186588.

2. Eligibility criteria

The search included studies that estimated sensitivity and specificity of ELISA or RDT index tests for chronic CD, with participants over five years old, patients with chronic CD, and patients without this disease; studies conducted in endemic and non-endemic areas for CD, that described the reference standards used, studies with a cross-sectional design and a case-control type; written in English, Spanish and Portuguese, published between 2010 and 2020; with research done with volunteers and with samples that included humans. Studies indicating that patients were receiving treatment for CD, those that were related exclusively to acute infection or in newborns, and those with mixed data on patients with acute and chronic infection were excluded.

3. Data sources

The databases used for the search, which was carried out from May to August 2020, were: Pubmed/Medline, Scopus; ISIWeb/Web of Science, and LILACS. The corresponding authors of articles included were contacted by email to inquire about missing data or request clarification on studies.

4. Study search and selection

The standard search strategy described in *The Joanna Briggs Institute Reviewers' Manual 2015* (16) was used. Thus, there was an initial limited search to identify relevant keywords and indexing terms, followed by a comprehensive search in the databases included with strategies for each of the search engines (S3). Two reviewers (SHSC-LXRL) assessed article titles and abstracts in an independent and blinded manner. Disagreements in the inclusion of studies were resolved by consensus, taking into account that the abstracts should meet the proposed eligibility criteria. Subsequently, the articles were reviewed in full text.

5. Data collection process

Two authors (SHSC-LXRL) extracted the following data independently: author(s), year of publication, type of participants, study area, index test, reference test, study period, country of implementation, number of patients and healthy subjects, total number of participants, sensitivity and specificity, risk of bias and applicability.

6. Definitions for data extraction

The subjects included in the different studies were classified into: patients who had lived or resided in an endemic area for CD and patients who reside in a non-endemic area.

The study area was considered endemic if CD occurred in this geographic area; and as a non-endemic area, otherwise. The index tests were considered commercial when they were part of a brand of laboratory diagnostic reagents, validated by medical device regulatory agencies and those available on the market; and considered in-house tests when studies indicated that immunoadsorption assays had been designed with different peptides or proteins with the application of non-standard "internal" methods. RDTs are those immunochromatographic assays that throw qualitative results and can be read at first sight.

Reference tests met the standard if they included a combination of serological tests with different antigens detecting antibodies against *T. cruzi,* and an additional test to reach a definitive diagnosis if the results were inconclusive.

The study design was considered clinical-comparative or case-control type if a group of participants diagnosed with chronic CD and a group without this diagnosis had been included; and it was considered non-comparative if a consecutive and representative series of patients with suspected CD had taken the test to be evaluated, as well as the reference test.

7. Risk of bias and applicability

Three authors (SHSC-LXLR-CSC) assessed the methodological quality and risk of bias of the studies included, in a blinded and independent manner, using the Quality Assessment of Diagnostic Accuracy Studies-2 (QUADAS-2) tool, which comprises four domains: patient

screening, index test, reference test, and flow and time (17). Each domain was assessed for risk of bias, and the first three domains were also assessed for applicability.

The QUADAS-2 tool was adjusted to the needs of this review, as follows: the risk of bias in patient screening was considered high if a consecutive or random sample of patients had not been used; and unclear if patient recruitment was not specified. The risk of bias related to the index test was considered unclear if there was no specification that the results of the index tests were interpreted without knowing the results of the reference test. The risk of bias related to reference tests was considered high if these tests were interpreted knowing the results of the index test, or if a single reference test had been used (taking into account that the WHO establishes that serological diagnosis in the chronic phase of CD should be based on positive results in two tests that are based on different immunological principles and, in case of inconsistency, on a third test).

8. Diagnostic accuracy measures

The reported measures were sensitivity and specificity for each of the index tests assessed for diagnosing chronic CD. When the studies did not have these two measures, they were calculated based on the number of true positives and negatives, as well as on the number of false positives and negatives and the total number of patients.

9. Summary of results

Sensitivity and specificity were modeled bivariately with binomial-normal random effects, with a gold standard (GS) assumption, but also with an imperfect gold standard (IGS) model.

The GS models were fitted with a Bayesian and classical approach; and the IGS model with a Bayesian approach only. Models were selected with the *deviance information criterion* (DIC) for the Bayesian models, and with the likelihood ratio test for the classical models. Six possible models for the GS were evaluated according to the type of distribution that followed the random effects (normal or mixed normal) and the type of connection (logit, cloglog and probit), and the best model was selected according to the smallest DIC with at least two points difference. The specification of the model with the best fit (in bamdit metadiag) was reproduced in the rest of the packages (meta4diag: Binomial-normal with probit, and metandi and IGS: Binomial-normal with logit) to facilitate comparisons.

The bivariate random effects model was used to estimate the overall sensitivity and specificity and their respective 95 % confidence intervals (CI). The results were plotted in *forest-plots* and ROC space (R DTAplots program), and heterogeneity between studies was assessed visually. R 1.3 *software* (DTAplots, bamdit::plotcompare and meta4diag::meta-regression) (18), Stata 15 (metandi) (19), midas and JAGS were used to conduct the meta-analysis.

10. Additional analyses

Meta-regressions were carried out with potential modifiers of diagnostic validity (bamdit plotcompare and meta4diag meta-regression). The variables of interest were study design (clinical comparative or non-comparative), study area (endemic or non-endemic), study risk (low or high risk of bias), sample type (serum, whole blood or not applicable) for the RDTs,

and type of test (commercial or in-house) for the ELISA tests but not for the RDTs because of the low number of studies, which made it impossible to estimate them.

All variables were categorized at two levels in both the ELISA and RDT assays to facilitate the comparison of predictive regions and validity estimates. A QUADAS-2 assessment was applied in each study in order to analyze by subgroups. The three levels of the QUADAS-2 became two: low risk and high risk (which included the *high risk and unclear* categories). Of the 7 items of the tool, item 1 (patient screening) and item 3 (reference standard) were considered since they were the only ones with a sufficient number of studies with a high risk of bias. In the rest of items, most studies were low risk.

A sensitivity analysis was carried out excluding influential outliers. Influential studies were reviewed based on the assumption that the subsequent interval distribution of study weight should include one. The publication bias was assessed using Deeks' asymmetry test, which was considered statistically significant with a value of $p < 0.1$ (20).