Is hyperpolarised gas magnetic resonance imaging a valid and reliable tool to detect lung health in cystic fibrosis patients? A COSMIN systematic review.

Fatmah Mallallah[1], Anna Packham[2], Ellen Lee[3], Daniel Hind[3].

1. Al-Adan Hospital, Ahmadi Governorate, Kuwait.
2. Sheffield Children's NHS Foundation Trust, Sheffield, UK.
3. Sheffield Clinical Trials Research Unit, Sheffield, UK.

Email addresses:

e.lee@sheffield.ac.uk
d.hind@sheffield.ac.uk

Abstract

This paper systematically reviewed the literature reporting the validity and reliability of hyperpolarised gas MRI as a marker of lung health in cystic fibrosis (CF). MEDLINE, EMBASE and grey literature were searched for studies assessing the measurement properties of hyperpolarised helium-3 or xenon-129 MRI. The COSMIN risk of bias tool was used to critically appraise eligible studies. Findings show hyperpolarised gas MRI was able to detect structural and functional abnormalities in the lungs, detect response to treatments, and is more sensitive than FEV_1 in detecting ventilation defects in CF patients. There was moderately robust evidence for construct validity of hyperpolarised gas MRI, although evidence for other types of validity is currently low. Nonetheless, high quality studies concluded that hyperpolarised gas MRI is a reliable tool and test results are reproducible in CF patients. Hyperpolarised gas MRI is a promising tool for detecting early CF pulmonary disease and for longitudinal monitoring of CF.

Introduction

Cystic Fibrosis[4] (CF) involves deteriorations in lung health due to the inability of the airways to clear accumulating mucus, making lungs more prone to respiratory tract infection and sputum production[1]. Spirometry, plethysmography, and multiple breath nitrogen washout (MBNW) (which measures lung clearance index (LCI)), are used in routine care to assess the severity of the disease and measure changes in lung volume. Computerized Tomography (CT), radiography and Magnetic Resonance Imaging (MRI) of the thorax can examine changes in lung structure, but cannot be used routinely to monitor the progression of the disease for safety and cost reasons[2,3].

Hyperpolarised (HP) gas magnetic resonance imaging (MRI) provides detailed resolution images by visualizing the distribution of the HP gas after inhalation [3–6]. Small areas of hypoventilation in the lungs give rise to a lower signal, quantified as ventilation defect percent (VDP). VDP can be quantified using different measurements

CF: Cystic Fibrosis, MBNW : Multiple Breath Nitrogen Washout , LCI: Lung Clearance Index, CT: Computerized Tomography, MRI: Magnetic Resonance Imaging, HP: Hyperpolarised, VDP: Ventilation Defect Percent, FEV_1 : Forced Expiratory Volume in 1 second, ^3He: Helium-3 gas, ^{129}Xe: Xenon-129, COSMIN: COnsensus-based Standards for the selection of health Measurement Instruments, PROMs: Patient-Reported Outcome Measures, ICC: Intra-Class coefficient, SDC: Smallest Detectable Change, LoA: Limits of Agreement, MIC: Minimal Important Change, GRADE: Grading of Recommendations Assessment, Development, and Evaluation, ANOVA: Analysis of Variance, SEM: Standard Error of Measurement, 95% CI: 95% Confidence Intervals, AUC: Area Under the Curve, PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses, RV/TLC: Residual Volume/Total Lung Capacity, RoB: Risk of Bias, CPT: Chest physiotherapy, ROC: Receiver operating characteristic.

[4]

such as k-means clustering, whole lung signal fraction, fuzzy c-means and linear binning[7]. VDP can be compared with other pulmonary function tests such as FEV_1 to get a better understanding of lung health[6]. Historically, the technique used helium-3 gas (^{3}He), but its relatively high cost and low availability has led to the increased use of xenon-129 (^{129}Xe). ^{129}Xe also dissolves more efficiently in the blood, providing better gas exchange information [6].

HP gas MRI has the potential to complement existing tests [4,5], but its current use is largely restricted to research purposes [8]. A systematic review of research on its measurement properties is needed to inform decisions about wider adoption. This paper aimed to systematically review primary research studies assessing the validity and reliability of HP ^{129}Xe or ^{3}He MRI as a marker of lung health.

Methods (1350)

The review was registered on PROSPERO database (CRD42019129588) before starting data extraction.

Eligibility criteria

Studies were eligible if they recruited people with CF aged 5 and over, irrespective of disease progression. Studies including patients with other conditions were included if the data for the CF group could be disaggregated. Eligible studies assessed the reliability and validity of HP ^{3}He or ^{129}Xe MRI. To be eligible a study had to report a summary statistic pertaining to at least one of the following: internal consistency; test-retest reliability; measurement error; content validity; construct validity; criterion validity [9,10]. FEV$_1$ was used as a criterion measure of validity – this being the gold standard measure of pulmonary function in clinical practice [11]. All studies aiming to develop or assess the measurement properties of HP gas MRI were included. There was no restriction on publication type; conference abstracts and theses were included. Studies using animal models and studies only assessing the feasibility or tolerability of HP gas MRI were excluded. Studies using HP gas MRI to validate another measure, or as an outcome were also excluded. Only papers published in the English language were included, due to resource constraints.

Systematic Literature Search

We searched MEDLINE and EMBASE via Ovid from inception to 21 August 2020, with no date restrictions, as well as EThOS for theses and Google Scholar for grey literature. We screened reference lists of eligible studies to identify further studies. Where necessary, we contacted authors to access unpublished data and identify further eligible studies. The combined thesaurus and free text terms related to the population and tests. The full electronic search strategy is on PROSPERO database (CRD42019129588). Two authors (FM, AP) independently screened titles and abstracts, then potentially relevant full-text articles for eligibility. Disagreements were resolved by a third reviewer (DH).

FM and AP extracted study characteristics (study objectives, design, sample size, age [however reported], and comparators) and summary statistics related to reliability, validity and responsiveness. For responsiveness, we assessed only ability of hyperpolarised gas MRI to detect changes in the lungs after treatment. We used the primary research studies' own hypotheses to assess construct validity, as no hypothesis had been set by the research team prior to data collection.

Risk of Bias Assessment

FM and AP assessed risk of bias using the COSMIN risk of bias checklist [12] (Table 1), with disagreements resolved by consensus. Although the COSMIN system was developed to assess the measurement properties of survey instruments (questionnaires), the underlying statistics used are the same as those used to evaluate the measurement properties of imaging techniques, and has sometimes been used for this purpose[13,14].

Rating the Evidence using COSMIN Criteria of Good Measurement Properties

The COSMIN criteria of good measurement were used to rate study results as sufficient, indeterminate and insufficient evidence of reliability or validity [15,16]. For test-retest reliability, an intra-class correlation (ICC) of 0.7 was rated sufficient; studies presenting no ICC were rated indeterminate. Sufficient evidence of an adequate measurement error required the smallest detectable change (SDC) or limits of agreement (LoA) to be less than the minimal important change (MIC). In the absence of the MIC, the findings were deemed insufficient evidence of measurement error. For criterion validity, correlation with the gold standard (FEV1) should be 0.7 or above to be rated sufficient. If this was not calculated, the study results would be rated indeterminate. For hypothesis testing (convergent validity) of construct validity, the results of the study should be in accordance with the study hypothesis to be rated sufficient. If no hypothesis was reported, the results would be rated indeterminate (Table 1).

Grading the Evidence using GRADE Approach

The overall quality of evidence was graded as very low (very little confidence in measurement property estimate), low (limited confidence in measurement property estimate), moderate (moderately confidence in measurement property estimate) or high (very confident that the measurement property estimate is close to the true measurement property estimate) using the Grading of Recommendations Assessment, Development, and Evaluation (GRADE) system [15,16]. Risk of bias, inconsistency, imprecision, and indirectness were used to determine the grade of the quality of evidence. Each measurement property begins on the High level, and may then be downgraded levels to moderate, low or very low as appropriate.

Risk of bias was assessed as: 1) no risk of bias (multiple studies with adequate risk of bias/at least one of very good quality); 2) serious (multiple studies of doubtful quality/at least one study with adequate quality); 3) very serious (multiple studies with inadequate risk of bias/at least one study with of doubtful quality); or, 4) extremely serious (one study with inadequate quality) (Table 1) [15,16].

Inconsistency was assessed as 1) acceptable (>75% study results in accordance), 2) serious (<75% study results in accordance), or 3) very serious (if all studies' results were insufficient). Imprecision refers to the total sample size and was assessed as 1) acceptable (n>100), 2) serious (n= 50 to 100), or 3) very serious (n<50). Indirectness refers to the study population including participants from other populations than the one of interest, and was assessed as 1) acceptable (only CF participants) 2) serious (healthy controls included in sample) and 3) very serious (This is not applicable to this study as only studies which included CF patients in the sample were included) (Table 1).

Summary Statistics Extracted

We extracted and summarised summary statistics. Reliability was measured by intraclass correlation (ICC), Bland–Altman, analysis of variance (ANOVA) and measurement error [9]. Measurement error was measured by standard error of measurement (SEM), Smallest Detectable Change (SDC) or Limits of Agreement (LoA) [9,10]. For the different types of validity: construct validity was measured by spearman rank correlation [10]; criterion validity was measured by standard correlation such as Pearson correlation, and area under the curve by calculating the sensitivity and specificity of the instruments used [10]; Responsiveness was assessed using 95% confidence intervals (95% CI), P-values and mean difference before and after the given treatment. This was assessing whether the HP gas MRI detected any changes in the lungs in response to the treatment.

Table 1: COSMIN Definitions and Methodology

Reliability/ validity measure	Definition	RoB Checklist	Good Measurement Properties	Grading Quality of Evidence (GRADE Approach)
Reliability				
Measurement error	The random and systematic error of a patient's result that is not associated to the true change in the construct to be measured. Measured by SEM, SDC or Limits of Agreement LoA [8,9]	· Patients stable in interim period · Time interval between scans appropriate · Test conditions similar for measurements · Measurement error: SEM, SDC of LoA calculated · Test retest: ICC calculated	+ SDC or LoA < MIC^5 ? MIC not defined – SDC or LoA > MIC^5	Number of levels to downgrade according to seriousness of each assessment: Risk of bias -0 Acceptable -1 Serious -2 Very serious -3 Extremely serious Inconsistency -0 Acceptable -1 Serious -2 Very serious Imprecision -0 Acceptable -1 total n=50-100 -2 total n<50 Indirectness -0 Acceptable -1 Serious -2 Very serious
Test- retest	The reproducibility of results if the test is repeated over time. Measured by intraclass correlation (ICC), Bland–Altman, analysis of variance (ANOVA) and measurement error [9].	· Patients stable in interim period · Time interval between scans appropriate · Test conditions similar for measurements · Measurement error: SEM, SDC of LoA calculated · Test retest: ICC calculated	+ ICC or weighted Kappa ≥ 0.70 ? ICC or weighted Kappa not reported - ICC or weighted Kappa < 0.70	
Validity		·		
Construct (convergent)	The degree to which the instrument relates to a measure it is hypothesised to have a strong relationship with. Measured by spearman rank correlation.	· Is it clear what the comparator (FEV$_1$) measures · Were measurement properties of comparator (FEV$_1$) sufficient · Design and statistical methods adequate for hypotheses to be tested	+ ≥ 75% study results in accordance with the study hypothesis ? No hypothesis defined by study authors - < 75% study results in accordance with the study hypothesis	
Criterion	The extent to which the results of an instrument reflect the gold standard measurement (FEV$_1$). Measured by standard correlation such as Pearson correlation, and AUC by calculating sensitivity and specificity of instruments used [9]	· Correlations or AUC calculated	+ Correlation with FEV$_1$ ≥ 0.70 OR AUC ≥ 0.70 ? Not all information for '+' reported - Correlation with FEV$_1$ < 0.70 OR AUC < 0.70	
Responsiveness	The extent to which an instrument is able to detect a clinically important change in the concept being measured.	· Correlations between change scores or AUC calculated	+ ≥ 75% study results in accordance with the hypothesis ? No hypothesis defined (by the review team) - The result is not in accordance with the hypothesis	

Table Note: Information in table taken from COSMIN Manual for Systematic Review of PROMs [12][15][16].

SEM = standard area of measurement, SDC = smallest detectable change, LoA = limits of agreement, AUC = area under receiver operator curve, ICC = intraclass correlation coefficient, MIC = minimal important change. + = sufficient, ? = indeterminate, - = insufficient.

Results

Following the PRISMA reporting guidelines[17], after the elimination of duplicates, the searches retrieved 204 articles through the electronic bibliographic databases and 10 citations through grey literature searching (total N=214) (Figure 1). After eligibility screening, 49 full-text articles were retrieved for eligibility assessment. Of these, 32 articles, representing 25 unique studies met the eligibility criteria and were included in the review (Table 2). Reasons for exclusion of studies at the full-text stage are given in Appendix 1.

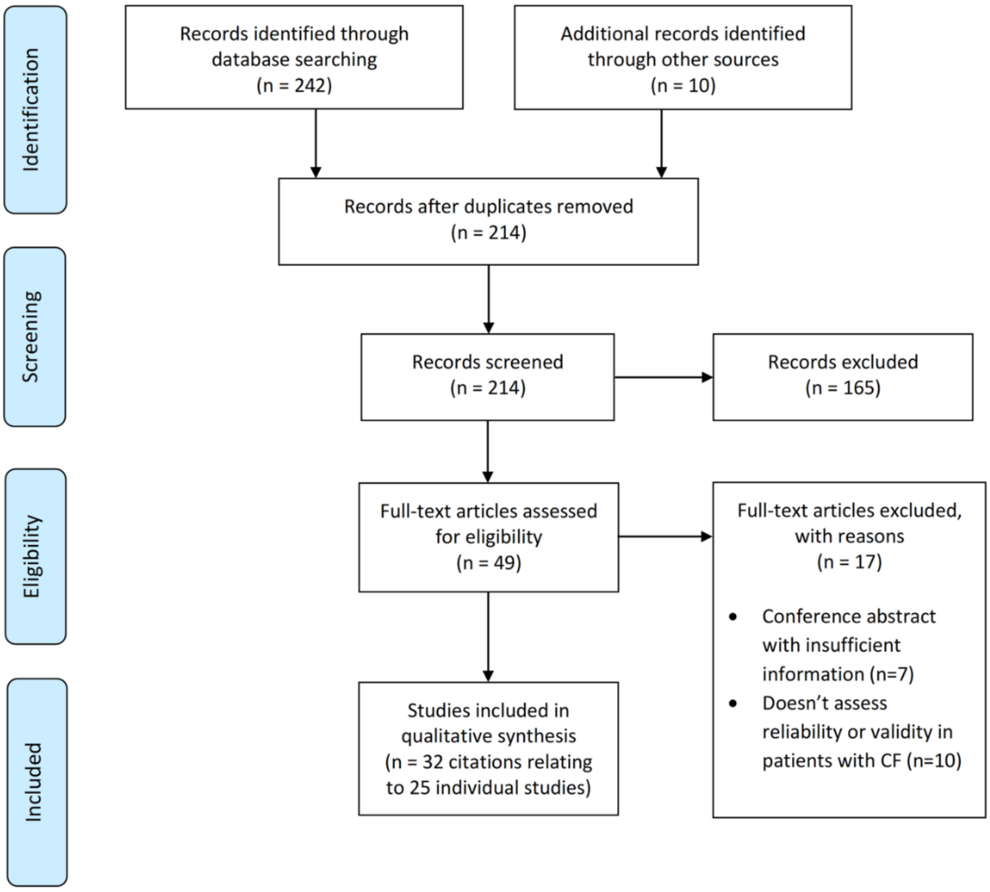
Identification
Records identified through database searching (n = 242)
Additional records identified through other sources (n = 10)
Records after duplicates removed (n = 214)
Screening
Records screened (n = 214)
Records excluded (n = 165)
Eligibility
Full-text articles assessed for eligibility (n = 49)
Full-text articles excluded, with reasons (n = 17)
• Conference abstract with insufficient information (n=7)
• Doesn't assess reliability or validity in patients with CF (n=10)
Included
Studies included in qualitative synthesis (n = 32 citations relating to 25 individual studies)

Figure 1: PRISMA 2009 Flow Diagram for Study Selection

Table 2: Study Characteristics by Validity/ Reliability Measure

Author (Country)	^{3}He or ^{129}Xe	Study Design	Study Duration	CF Sample Size	CF Age Mean/Median (SD/Range)	Comparator Device
Measurement error						
Kirby et al, 2011, Canada [18]	^{3}He	Case series	2 scans in 1 week	12	Mean = 26 (range 18 to 41)	Spirometry; plethysmography
Test-retest reliability						
Woodhouse et al, 2009, UK [19]	^{3}He	Cross-sectional	2 scans in 1 session	5	Mean = 11 (range 6 to 15)	Spirometry
Choy et al, 2010, Canada [20][21]	^{3}He	Pilot study	2 scans in 1 week	8	Mean = 25 (SD=8)	Spirometry
Bannier et al, 2010, France [22]	^{3}He	Cross-sectional	2 scans in 1 session	10	Mean = 10.2 (range 8 to 16)	Spirometry (CPT was done for all patients to check changes in HP MRI)
O'Sullivan et al, 2014, Canada [23]	^{3}He	Case series	4 scans in 4 weeks	5	NA	Spirometry
Zha et al, 2019, USA [8]	^{3}He	Cross-sectional and Case series study	2 scans in 2 weeks	7	Mean = 23.8 (SD=10.5)	Spirometry
Couch et al, 2019, Canada and USA [24]	^{129}Xe	Retrospective analysis	2 scans	CF = 18 HC = 8	CF Mean = 13.1 (SD=2.3) Healthy Mean = 12.7 (SD=2.3)	Spirometry, plethysmography; MBNW
Smith et al,2020, UK [25]	^{129}Xe	Case series	1 scan at baseline and at 16 month follow up (n=18) 2 scans at baseline and at 16 month follow up (n=11)	29	Mean = 23.0 (SD=11.1)	spirometry, plethysmography; MBNW
Smith et al, 2019, UK [26,27]	^{3}He and ^{129}Xe	Cross-sectional and Longitudinal study	2 scans in 20 months	14	Median 17.4 (range 6.4 – 47.5)	Spirometry; MBNW
Criterion validity						
Koumellis et al, 2005, USA [28]	^{3}He	Cross-sectional	1 scan	8	Mean = 11.4 (range 6 to 15)	Spirometry

Mentore et al, 2005, USA [29]	^{3}He	Case series	3 scans for 3 different treatments	CF= 15 HC = 16	CF Mean = 21 (range 15 to 33) Healthy Mean = 25 (range 21 to 33)	Spirometry
Van Beek et al, 2006, UK [30]	^{3}He	Cross-sectional	1 scan	18	Median 12.1 (range 5 to 17)	Spirometry; chest X-ray
Woodhouse et al, 2009, UK [19]	^{3}He	Cross-sectional	2 scans in 1 session	14	Reproducibility cohort: Mean = 11 (range 6 to 15) Intervention cohort: Mean = 9 (range 5 to 15)	Spirometry
Choy et al, 2010, Canada [20][21]	^{3}He	Pilot study	2 scans in 1 week	8	Mean 25 (SD=8)	Spirometry
Kirby et al, 2013, Canada [31]	^{3}He	Case series	2 scans in 1 week	11	Mean 27 (SD=8)	Spirometry; plethysmography
Paulin et al, 2015, Canada [32]	^{3}He	Case series	3 scans in 4 years	5	28 (range 20 to 36)	Spirometry; plethysmography
Hardy et al, 2016, UK [33]	^{3}He	Cross-sectional	1 scan	CF= 18 HC = 30	CF Mean = 14.7 (SD=2.3) Healthy Mean = 14.6 (SD=1.4)	Spirometry; plethysmography; MBNW
Marshall et al, 2017, UK [34]	^{3}He	Cross-sectional	1 scan	CF = 19 HC = 10	CF Mean = 10.9 (SD=2.5) Healthy Mean= 11.3 (SD=2.8)	Spirometry; plethysmography; MBNW
Smith et al, 2019, UK [35]	^{3}He	Case series	Scan at baseline and at 16 month follow up	28	NA	Spirometry; MBNW
Thomen et al, 2017, USA[36]	^{129}Xe	Cross-sectional	1 scan	CF = 11 HC = 11	CF Mean = 12.5 (SD=2.3) Healthy Mean = 11.5 (SD=3.2)	Spirometry
Kanhere et al, 2017, Canada [37]	^{129}Xe	Cross-sectional	1 scan	CF = 10 HC = 5	CF Mean = 13 (SD=2.5) Healthy Mean = 12.4 (SD=2.4)	Spirometry; plethysmography; MBNW

Couch et al, 2019, Canada and USA [24]	^{129}Xe	Retrospective analysis	1 scan	CF = 18 HC = 8	CF Mean = 13.1(SD=2.3) Healthy Mean = 12.7 (SD=2.3)	spirometry, plethysmography; MBNW
Rayment et al, 2019, Canada [38,39]	^{129}Xe	Cohort study	2 scans pre and post treatment over 3 weeks	15	Median of 14 (range 13.0 to 16.5)	Spirometry, plethysmography;MBNW
Construct validity						
McMahon et al, 2006, Ireland [4]	^{3}He	Case series	1 scan	8	Mean = 31.9 (range 20 to 46)	HRCT; spirometry
Bannier et al, 2010, France [22]	^{3}He	Cross-sectional	2 scans in 1 session	10	Mean = 10.2 (range 8 to 16)	Spirometry (CPT was done for all patients to check changes in HP MRI)
Kirby et al, 2011, Canada [18]	^{3}He	Case series	2 scans in 1 week	12	Mean = 26 (range 18 to 41)	Spirometry; plethysmography
Altes et al, 2012, USA [40–42]	^{3}He	Study A: Crossover clinical trial Study B: open label trial	5 scans in 48 weeks	Study A = 8 Study B = 9	A Mean = 18.9 B Mean = 24.4	Spirometry; MBNW
Smith et al, 2018, UK [43][44]	^{3}He	Cross-sectional	1 scan	32	Median 16.7 (range 6.4–43.1)	Spirometry; plethysmography; MBNW
Smith et al, 2018, UK [45]	^{3}He	Case series	2 scans in 1.3-2 years	14	Mean = 10.30 (SD=2.26)	Spirometry; plethysmography; MBNW
Zha et al, 2019, USA [8]	^{3}He	Cross-sectional and Case series study	1 scan	17	Mean = 23.8 (SD=10.5)	Spirometry
Smith et al, 2019, UK [26,27]	^{3}He and ^{129}Xe	Cross-sectional and Longitudinal study	One (n=17) to two (n=14) scans in 20 months	31	Median 17.4 (range 6.4 – 47.5)	Spirometry; MBNW
Responsiveness						
Bannier et al, 2010, France [22]	^{3}He	Cross-sectional	2 scans (pre- and post-CPT)	10	Mean = 10.2 (range 8 to 16)	Spirometry (CPT was done for all patients to check changes in HP MRI)
Altes et al, 2012, USA [40–42]	^{3}He	Study A: Crossover clinical trial Study B: open label trial	5 scans in 48 weeks	Study A = 8 Study B = 9	A Mean = 18.9 B Mean = 24.4	Spirometry; MBNW
Rayment et al, 2019, Canada [38,39]	^{129}Xe	Cohort study	2 scans (pre- and post-treatment) over 3 weeks	15	Median of 14 (range 13.0 to 16.5)	Spirometry, plethysmography; MBNW

Smith et al,2020, UK [46]	^{129}Xe	Cross-sectional	2 scans (pre- and post-exercise)	13	Mean = 25 (SD=10)	Spirometry, plethysmography; MBNW
Woodhouse et al, 2009, UK [19]	^{3}He	Cross-sectional	2 scans (pre- and post-physiotherapy)	9	Mean = 9 (range 5 to 15)	Spirometry
Mentore et al, 2005, USA [29]	^{3}He	Case series	3 scans for 3 different treatments	CF= 15 HC = 16	CF Mean = 21 (range 15 to 33) Healthy Mean = 25 (range 21 to 33)	Spirometry

CPT= Chest physiotherapy, MBNW = Multiple breath nitrogen washout

Study Characteristics

The 25 included studies were published between 2005 and 2020. Six were conducted in the USA [8,24,28,29,36,40], seven in Canada [18,20,23,31,32,37,38], 10 in the UK [19,25,26,30,33–35,43,45,46], one in France [22], and one in Ireland [4]. Eighteen of the studies investigated HP ^{3}He MRI [4,8,18–20,22,23,28–35,40,43,45] six investigated HP ^{129}Xe MRI [24,25,36–38,46] and one investigated both HP ^{3}He and ^{129}Xe MRI [26] (Table 2).

There were 11 cross-sectional studies [8,19,22,25,28,30,33,34,36,37,43], six case series with a follow-up of less than twelve months [4,18,23,29,31] and four [25,32,35,45] with follow-up of between 14 months [25] and 4 years [32]. Two case series exposed participants to interventions such as nebulisers and chest physiotherapy prior to the MRI scan to understand treatment response [18,29]. One study presented a nested case series within a larger cross-sectional study [26]. There was one crossover clinical trial [40], There was one pilot study [20], one retrospective analysis study [24], and one cohort study [38].

Sample sizes ranged from five [23,32] to 32 [43] people with CF and from 5 [37] to 30 [33] healthy individuals, in studies which used controls. The reported mean age of study populations ranged from 9 to 32. Reported median ages ranged from 12.1 to 17.4.

Test-retest Reliability

Eight studies assessed test-retest reliability [8,19,20,22–26]; five using the ICC [8,20,22,24,25], four using Bland-Altman tests [8,19,25,26], and one using ANOVA [23] (Table 4). There was good evidence for the test-retest reliability of MRI in assessing VDP across three studies of very good [25] and adequate quality [20,22] in which the intraclass correlations were more than 0.9 (Table 4). The GRADE assessment was very low for studies using HP ^{3}He as there was inconsistency between studies, with <75% of the studies showing a strong correlation, and considerable imprecision (total n=49 study participants with cystic fibrosis) (Table 6). For studies using HP ^{129}Xe the GRADE assessment was low due to Risk of Bias (only two studies of doubtful and inadequate quality), imprecision (total n=61 study participants with cystic fibrosis) and inconsistency with <75% of the studies showing a strong correlation.

Measurement Error

There was poor evidence for measurement error from a single study of adequate quality [18] (Table 5). This study found the SDC for VDP (3%) to be higher than the MIC (2%), suggesting a considerable chance that the change detected by MRI for VDP was caused by measurement error [10,47]. The GRADE assessment was very low due to risk of bias (only one study of adequate quality), imprecision (total n=12 study participants with cystic fibrosis) and difficulty in assessing consistency.

Criterion Validity

From the HP ^{3}He and ^{129}Xe MRI studies, fourteen assessed criterion validity [19,20,24,28–38] (Table 6). One study [28] presented p-values, which were not interpretable, rather than correlations. Evidence for criterion validity was mixed when using FEV1 as a criterion of VDP. Four studies of very good quality showed a strong correlation (>0.7) between FEV_1 and VDP [19,29,32,34], however six very good quality studies[20,24,31,33,36,38] found a weaker correlation (<0.7, range: 0.3-0.69). The GRADE assessment for criterion validity was low for studies using HP ^{3}He, and very low for studies using HP ^{129}Xe. These grades are due to inconsistency (<75% showed strong correlation) and indirectness (total healthy controls: ^{129}Xe n=24; ^{3}He n=56), with HP ^{129}Xe MRI being downgraded further for imprecision (total n=54 study participants with cystic fibrosis).

Of the three studies of very good quality which assessed LCI as a criterion for VDP, one study [24] showed a strong correlation (>0.7) and two [33,38] a weaker correlation (<0.7, range: 0.13-0.61). Low correlations were found between HP gas MRI and body plethysmography; RV/TLC [34,38]; and CT scan score [34] (Table 6).

Construct Validity

Seven studies using HP ^{3}He MRI assessed construct validity [4,8,18,22,40,43,45], in addition to one study which used both HP ^{3}He and ^{129}Xe [26] (Table 7). Four of the eight studies did not report a study hypothesis and were rated indeterminate [22,26,43,45] (Appendix 2). From the four studies which included a hypothesis, two studies of very good quality [4,40] found a strong correlation between FEV_1 and VDP (>0.7) in accordance with their hypothesis, and one study of very good quality found a weaker correlation (= -0.68) [18]. The GRADE assessment for construct validity was moderate for the studies using HP ^{3}He, after being downgraded for inconsistency, with less than 75% studies having sufficient results. The GRADE assessment for construct validity of studies using HP ^{129}Xe was very low, due to difficulty assessing inconsistency of evidence from only one study, and imprecision (total n=31 patients with cystic fibrosis).

From the studies of very good quality which assessed construct validity against other techniques, strong correlations (>0.7) were found between hyperpolarised gas MRI and LCI in two studies [26,43], RV/TLC in one study [43], and CT scan in one study [4].

Responsiveness

Six studies assessed responsiveness to treatment [19,22,29,38,40,46]. There were two studies [29,40] of very good quality which found HP 3He MRI was able to detect changes in ventilation volume and defects after treatment, however the evidence was rated as intermediate due to no hypothesis being set by the review team (Table 8).

Table 4: Test-Retest Reliability: Risk of Bias Within Studies, Good measurement properties according to COSMIN Checklist, Study Findings, and GRADE result

Authors (Date)	^{3}He or ^{129}Xe	RoB	Good Measurement Properties	Study Findings	Number of patients in all of the studies	GRADE
Woodhouse et al (2009) [19]	^{3}He	Doubtful	Indeterminate	Bland–Altman analysis for both examinations, The mean difference between the two examinations = -0.037 (95% CI -7.7 to 0.15)	(n=49)	Very Low
Choy et al (2010) [20]	^{3}He	Adequate	Sufficient	ICC of Ventilation Gradients (VG3x3) of VDP = 0.92 ICC of Coefficients of Variation (CoV3x3) of VDP = 0.91		
Bannier et al (2010) [22]	^{3}He	Adequate	Sufficient	ICC of VDP = 0.924		
O'Sullivan et al (2014) [23]	^{3}He	Inadequate	Indeterminate	ANOVA of VDP P= 0.2871		
Zha et al (2019) [8]	^{3}He	Doubtful	Sufficient	Bland-Altman analysis of VDP = 0.023 (95% CI -0.06 to 0.105) ICC of VDP = 0.95		
Smith et al, (2019) [26]	^{3}He	Doubtful	Indeterminate	Bland–Altman analysis of ^{3}He-^{129}Xe HP MRI LoA = 8.9, -7.4%		
Couch et al (2019) [24]	^{129}Xe	Inadequate	Sufficient	ICC of VDP = 0.99	(n=61)	Low
Smith et al, (2020) [25]	^{129}Xe	Very good	Sufficient	Bland-Altman analysis (LoA) of VDP = 0.8 [-7.0, 8.5] ICC of VDP = 0.97 [0.94, 0.99]		

Smith et al, (2019) [26]	^{129}Xe	Doubtful	Indeterminate	Bland–Altman analysis of ^{3}He-^{129}Xe HP MRI LoA = 8.9, -7.4%		

CI= Confidence intervals, ICC= Intraclass correlation coefficient, VDP= Ventilation defect percentage

Table 5: Measurement Error: Risk of Bias Within Studies, Good measurement properties according to COSMIN Checklist, Study Findings, and GRADE result

Authors (Date)	^{3}He or ^{129}Xe	RoB	Good Measurement Properties	Study Findings	Number of patients in all of the studies	GRADE
Kirby et al (2011) [18]	^{3}He	Adequate	Insufficient	SDC in VDP = 0.03	(n=12)	Very Low

SDC= Smallest detectable change, VDP= Ventilation defect percentage

Table 6: Criterion Validity: Risk of Bias Within Studies, Good measurement properties according to COSMIN Checklist, Study Findings, and GRADE result

Authors (Date)	^{3}He or ^{129}Xe	RoB	Good Measurement Properties	Study Findings	Number of patients in all of the studies	GRADE
Koumellis et al (2005) [28]	^{3}He	Inadequate	Indeterminate	Correlation between VDP and FEV1 not reported. The six peripheral ROI measurements were averaged to obtain an index of flow in the peripheral lung, a good correlation with the FEV_1 analysed by means of a two-tailed Student's t-test was found P= 3.74 x 10^{-5}	(n=144)	Low
Mentore et al (2005) [29]	^{3}He	Very Good	Sufficient	Correlation coefficients (r) for: · VDP with FEV_1 = - 0.71		
Beek et al (2006) [30]	^{3}He	Doubtful	Insufficient	Correlation coefficients (r) for: · ^{3}He MRI with FEV_1 = - 0.41		

Woodhouse et al (2009) [19]	^{3}He	Very Good	Sufficient	Correlation coefficients (r) for: · ^{3}He MRI with FEV_1 = 0.98 · ^{3}He MRI with FEV_1 = 0.82
Choy et al (2010) [20]	^{3}He	Very Good	Insufficient	Correlation coefficients (r) for: · Ventilation Gradients (VG3x3) of VDP with FEV_1 =0.69 · Coefficients of Variation (CoV3x3) of VDP with FEV_1 = 0.66
Kirby et al (2013) [31]	^{3}He	Very Good	Sufficient	Correlation coefficients (r) for: · FEV_1 with difference in whole lung apparent diffusion coefficient (ADC) (^{3}He MRI) = 0.67 · FEV_1 with previously ventilated ADC interior posterior difference = -0.75
Paulin et al (2015) [32]	^{3}He	Very Good	Sufficient	Linear regression (r2) for: · Baseline VDP with FEV_1 after 4 years = 0.98 · 4-year VDP with FEV_1 after 4 years = 0.85
Hardy et al (2016) [33]	^{3}He	Very Good	Insufficient	Correlation between VDP and FEV1 not reported. Correlation coefficients (r) for: · 13 ms ADC (VC W) with FEV_1 = - 0.39

				· 13 ms ADC (VC W) with LCI = -0.13		
Marshall et al (2017) [34]	^{3}He	Inadequate	Sufficient	AUC for: · ^{3}He MRI VDP = 0. 94 Correlation coefficients (r) for: · VDP with RV/TLC = 0.61 · VDP with CT gas trapping score = 0.58 · VDP with LCI siting = 0.55		
Smith et al, (2019) [35]	^{3}He	Inadequate	Indeterminate	Correlation between VDP and FEV1 not reported Correlation coefficients (r) for: · The difference in VDP from baseline to the follow-up with the difference in LCI from baseline to the follow-up = 0.61		
Kanhere et al (2017) [37]	^{129}Xe	Doubtful	Insufficient	Coefficient of multiple correlation (r^2): · VDP with FEV_1 in all patients (CF and HC) = 0.31 · VDP with LCI in all patients (CF and HC) = 0.88	(n=54)	Very Low
Thomen et al (2017) [36]	^{129}Xe	Very Good	Insufficient	Correlation coefficients (r) for: · VDP with FEV_1 = - 0.54		
Couch et al (2019) [24]	^{129}Xe	Very Good	Insufficient	Linear regression (r2) for:		

				· VDP with FEV_1 done by analyst 1 = 0.33 · VDP with FEV_1 done by analyst 2 = 0.26 · VDP with LCI done by analyst 1 = 0.76 · VDP with LCI done by analyst 2 = 0.77
Rayment et al (2019) [38]	^{129}Xe	Very Good	Insufficient	Linear regression (r2) for: · VDP with FEV_1 = 0.30 · VDP with LCI supine = 0.21 · VDP with LCI seated = 0.38 · VDP with RV/TLC = 0.34

VDP= Ventilation defect percentage, FEV_1 = Forced expiratory volume in 1 second, ROI= Regions of interest, ADC = apparent diffusion coefficient, VCW= Weighted and volume corrected, RV= Residual volume, TLC= total lung capacity, AUC= Area under curve, CT= Chest tomography, LCI= lung clearance index, HC = healthy controls

Table 7: Construct Validity: Risk of Bias Within Studies, Good measurement properties according to COSMIN Checklist, Study Findings, and GRADE result

Authors (Date)	^{3}He or ^{129}Xe	RoB	Good Measurement Properties	Study Findings	Number of patients in all of the studies	GRADE
McMahon et al (2006) [4]	^{3}He	Very Good	Sufficient	Spearman rank correlation (ρ) for: · FEV_1 with VDP = 0.86 · HRCT with VDP = ± 0.89	(n=148)	Moderate
Bannier et al (2010) [22]	^{3}He	Very Good	Insufficient	Spearman rank correlation (ρ) for: · FEV_1 with VDP = - 0.041		
Kirby et al (2011) [18]	^{3}He	Very Good	Insufficient	Spearman rank correlation (ρ) for:		

				· FEV_1 with VDP = - 0.68		
Altes et al (2012) [40]	^{3}He	Very Good	Sufficient	Spearman rank correlation (ρ) for: · FEV_1 with VDP = - 0.85 · Part A of the study, Spearman rank correlation (ρ) for: ▪ FEV_1 with VDP = - 0.52 · Part B of the study, Spearman rank correlation (ρ) for: ● FEV_1 with VDP = - 0.67		
Smith et al (2018) [43]	^{3}He	Very Good	Indeterminate	Spearman rank correlation (ρ) for: · VDP with FEV_1 = - 0.79 · VDP with LCI = 0.89 · VDP with RV/TLC = 0.80		
Smith et al (2018) [45]	^{3}He	Doubtful	Indeterminate	Correlation between VDP and FEV_1 not reported Spearman rank correlation (ρ) for: · VDP and LCI at baseline = 0.66 · VDP and LCI at visit 2 = 0.82 · The percentage change in VDP from baseline to visit 2 = 0.60		
Zha et al (2019) [8]	^{3}He	Doubtful	Sufficient	Spearman rank correlation (ρ) for: · VDP with FEV_1 = - 0.75		
Smith et al, (2019) [26]	^{3}He	Very Good	Indeterminate	Spearman rank correlation (ρ) for: · VDP with FEV_1 = - 0.78 · VDP with LCI = 0.88		
Smith et al, (2019) [26]	^{129}Xe	Very Good	Indeterminate	Spearman rank correlation (ρ) for:	(n=31)	Very Low

- VDP with FEV_1 = - 0.79
- VDP with LCI = 0.88

FEV_1 = Forced expiratory volume in 1 second, VDP= Ventilation defect percentage, LCI= Lung clearance index, RV= Residual volume, TLC= total lung capacity

Table 8: Responsiveness: Risk of Bias Within Studies, Good measurement properties according to COSMIN Checklist, Study Findings, and GRADE result

Authors (Date)	^{3}He or ^{129}Xe	RoB	Good Measurement Properties		GRADE
Mentore et al (2005) [29]	^{3}He	Very Good	Indeterminate	Correlation between change in VDP and FEV1 (% predicted) from baseline to after albuterol, DNase, and chest physical therapy was weak (r = –0.13).	N/A
Woodhouse et al (2009) [19]	^{3}He	Inadequate	Indeterminate	The was no significant changes in total ventilation volume before and after CPT that was detected using hyperpolarised helium MRI (P value = 0.36)	
Bannier et al (2010) [22]	^{3}He	Inadequate	Indeterminate	VDP before and after CPT did not change (P value > 0.10); · VDP mean before CPT = 5.1 (1.9) · VDP mean after CPT = 5.1 (1.1)	
Altes et al (2012) [40]	^{3}He	Very Good	Indeterminate	Part A: VDP was reduced by 8.2% from baseline (day 15) to after ivacaftor treatment (day 43), P value = 0.055 (r= −0.5238)	
Rayment et al (2019) [38]	^{129}Xe	Inadequate	Indeterminate	The absolute mean change in VDP pre- post treatment was -3.0 (-5.0, -1.0) The relative change (%) in VDP pre- post treatment was -44.2 (-60.2, -28.3)	N/A
Smith et al, (2020) [46]	^{129}Xe	Inadequate	Indeterminate	There was a small but significant reduction in the VDP (p = 0.04) after CPET when compared to baseline. · VDP % before CPET = 7.3 [2.3, 25.8] · VDP % after CPET = 7.1 [2.4, 24.8]	

VDP= Ventilation defect percentage, FEV_1 = Forced expiratory volume in 1 second, CPT= Chest physiotherapy, CPET= Cardiopulmonary exercise testing

Summary of Risk of Bias Issues

The greatest risk of bias for tests of reliability resulted from inadequate reporting of patient stability on the day of scans [8,18,19,21,22,26] and consistency in test conditions for all participants [8,18,19,26]. Reporting of the summary statistics necessary to understand measurement properties was often inadequate. For instance, ICCs, mean differences and 95% CIs were often not reported for correlations [23,24,30,37]. Risk of bias arose in studies of construct and criterion validity which did not correlate HP gas MRI against FEV_1 (the current gold standard) [8,28,34,35,45]. Further to this, studies of responsiveness frequently failed to correlate changes in HP gas MRI with those observed in FEV_1 [19,22,38,46].

Summary of Issues arising during Grading

Evidence for reliability was frequently downgraded for inconsistency in study results and imprecision due to the studies' small sample size. Evidence for validity was downgraded for inconsistency in findings, indirectness due to heterogeneity in samples including healthy participants, and imprecision arising from low statistical power.

Discussion

This review found moderately robust evidence for construct validity of HP gas MRI as a marker of lung health in people with cystic fibrosis. Evidence for other types of validity and reliability is currently low. Nonetheless, high quality studies [4,20,22,27,43] concluded that HP gas MRI was a useful tool to detect lung ventilation defects, was useful in tracing the functional and structural progression of cystic fibrosis, and test results were reproducible in cystic fibrosis patients.

HP gas MRI was able to detect ventilation defects in patients with normal FEV_1 results [22,24,31,34,36] and is better able to discriminate CF patients from healthy controls than FEV_1 [36], especially in children where the disease is still developing. While FEV_1 was sensitive in detecting obstruction in large airways, FEV_1 cannot detect ventilation defects in small airways [48,49]. That HP gas MRI can detect changes in VDP over short periods of time, indicates its potential in the management of CF [18,31].

It is important to note that there is currently a lack of standardisation in the acquisition and analysis of HP gas MRI. This is a limitation of this review, as the differences in the generation of VDP could mean the measures of validity and reliability are not be directly comparable[50][51].

Future research needs to serve both a policy-making and clinical audience, including those who still see FEV_1 as the gold standard and those who see it as an insensitive measure of lung health. Advocates of HP gas MRI have made the case, successfully, that it detects functional defects in CF patients better than other methods. To bring about a shift in clinical and policy norms requires an argument about why that matters in clinical terms and is cost-effective for health systems. In particular, overuse of imaging has adverse economic consequences and is burdensome for patients in terms of repeat tests and exposure to x-rays [52]. Qualitative research is needed to assess the degree to which cystic fibrosis specialists (respiratory physicians and physiotherapists) consider HP gas MRI an adequate reflection of lung health (content validity). Given the long natural history of cystic fibrosis, decision-makers may require that changes in gas MRI-assessed lung health are validated against FEV_1 and other instruments, such as LCI, over a period of at least five years to demonstrate MRI's prognostic validity and clinical potential.

To reduce risk of bias, future studies must document patient stability on the day of testing, and report consistency in test conditions, for all participants. Authors of studies which reported patient stability defined this as no changes in respiratory symptoms or medications in the period leading up to participation in the study, which ranged from 1 week to 4 weeks [20,23–25]. To improve the statistical quality of studies, when analysing correlations, the ICC, mean difference and 95% CI should be reported.

Conclusion

HP gas MRI is a promising tool for detecting early CF pulmonary disease and for longitudinal monitoring of the progression of the disease. It is more sensitive than FEV_1, in detecting functional and structural ventilation defects in CF patients and is responsive to CF pulmonary treatments. Further validation is required against a range of measures in long-term studies to assess its prognostic value and cost-effectiveness.

Acknowledgements.

No specific funding supported this systematic review. DH conceived of the study. DH and FM designed the searches and, with AP, drafted the manuscript. DH and FM

designed the study. FM and AP ran searches, screened studies for eligibility, extracted data and critically appraised primary research studies. EL was the study statistician and contributed to the first and subsequent drafts of the manuscript. DH, AP, FM and EL commented on and approved the final manuscript.

Conflicts of Interests.

All authors declare that they have no competing interests.

Reference List

[1] Davies JC, Alton EWFW, Bush A. Cystic fibrosis. BMJ 2007;335:1255–9. https://doi.org/10.1136/bmj.39391.713229.AD.

[2] Mayo JR, Aldrich J, Müller NL. Radiation Exposure at Chest CT: A Statement of the Fleischner Society. Radiology 2003;228:15–21. https://doi.org/10.1148/radiol.2281020874.

[3] Donnelly LF, MacFall JR, McAdams HP, Majure JM, Smith J, Frush DP, et al. Cystic Fibrosis: Combined Hyperpolarized 3 He-enhanced and Conventional Proton MR Imaging in the Lung—Preliminary Observations. Radiology 1999;212:885–9. https://doi.org/10.1148/radiology.212.3.r99se20885.

[4] McMahon CJ, Dodd JD, Hill C, Woodhouse N, Wild JM, Fichele S, et al. Hyperpolarized 3helium magnetic resonance ventilation imaging of the lung in cystic fibrosis: comparison with high resolution CT and spirometry. Eur Radiol 2006;16:2483–90. https://doi.org/10.1007/s00330-006-0311-5.

[5] Evans A, McCormack DG, Santyr G, Parraga G. Mapping and quantifying hyperpolarized 3 He magnetic resonance imaging apparent diffusion coefficient gradients. J Appl Physiol 2008;105:693–9. https://doi.org/10.1152/japplphysiol.00178.2008.

[6] Santyr G, Kanhere N, Morgado F, Rayment JH, Ratjen F, Couch MJ. Hyperpolarized Gas Magnetic Resonance Imaging of Pediatric Cystic Fibrosis Lung Disease. Acad Radiol 2019;26:344–54. https://doi.org/10.1016/j.acra.2018.04.024.

[7] Woodhouse N, Wild JM, Paley MNJ, Fichele S, Said Z, Swift AJ, et al.

Combined helium-3/proton magnetic resonance imaging measurement of ventilated lung volumes in smokers compared to never-smokers. J Magn Reson Imaging 2005;21:365–9. https://doi.org/10.1002/jmri.20290.

[8] Zha W, Nagle SK, Cadman R V., Schiebler ML, Fain SB. Three-dimensional Isotropic Functional Imaging of Cystic Fibrosis Using Oxygen-enhanced MRI: Comparison with Hyperpolarized 3 He MRI. Radiology 2019;290:229–37. https://doi.org/10.1148/radiol.2018181148.

[9] Bruton A, Conway JH, Holgate ST. Reliability: What is it, and how is it measured? Physiotherapy 2000;86:94–9. https://doi.org/10.1016/S0031-9406(05)61211-4.

[10] Scholtes VA, Terwee CB, Poolman RW. What makes a measurement instrument valid and reliable? Injury 2011;42:236–40. https://doi.org/10.1016/j.injury.2010.11.042.

[11] Barreiro TJ, Perillo I. An approach to interpreting spirometry. Am Fam Physician 2004;69:1107–14.

[12] Mokkink LB, de Vet HCW, Prinsen CAC, Patrick DL, Alonso J, Bouter LM, et al. COSMIN Risk of Bias checklist for systematic reviews of Patient-Reported Outcome Measures. Qual Life Res 2018;27:1171–9. https://doi.org/10.1007/s11136-017-1765-4.

[13] Saltzherr MS, Selles RW, Bierma-Zeinstra SMA, Muradin GSR, Coert JH, van Neck JW, et al. Metric properties of advanced imaging methods in osteoarthritis of the hand: a systematic review. Ann Rheum Dis 2014;73:365–75. https://doi.org/10.1136/annrheumdis-2012-202515.

[14] Jaspers MEH, van Haasterecht L, van Zuijlen PPM, Mokkink LB. A systematic

review on the quality of measurement techniques for the assessment of burn wound depth or healing potential. Burns 2019;45:261–81. https://doi.org/10.1016/j.burns.2018.05.015.

[15] Terwee CB, Prinsen CAC, Chiarotto A, Westerman MJ, Patrick DL, Alonso J, et al. COSMIN methodology for evaluating the content validity of patient-reported outcome measures: a Delphi study. Qual Life Res 2018;27:1159–70. https://doi.org/10.1007/s11136-018-1829-0.

[16] Prinsen CAC, Mokkink LB, Bouter LM, Alonso J, Patrick DL, de Vet HCW, et al. COSMIN guideline for systematic reviews of patient-reported outcome measures. Qual Life Res 2018;27:1147–57. https://doi.org/10.1007/s11136-018-1798-3.

[17] Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Med 2009;6:e1000097. https://doi.org/10.1371/journal.pmed.1000097.

[18] Kirby M, Svenningsen S, Ahmed H, Wheatley A, Etemad-Rezai R, Paterson NAM, et al. Quantitative Evaluation of Hyperpolarized Helium-3 Magnetic Resonance Imaging of Lung Function Variability in Cystic Fibrosis. Acad Radiol 2011;18:1006–13. https://doi.org/10.1016/j.acra.2011.03.005.

[19] Woodhouse N, Wild JM, van Beek EJR, Hoggard N, Barker N, Taylor CJ. Assessment of hyperpolarized 3 He lung MRI for regional evaluation of interventional therapy: A pilot study in pediatric cystic fibrosis. J Magn Reson Imaging 2009;30:981–8. https://doi.org/10.1002/jmri.21949.

[20] Choy S, Ahmed H, Wheatley A, McCormack DG, Parraga G. Development of spatial-temporal ventilation heterogeneity and probability analysis tools for

hyperpolarized 3 He magnetic resonance imaging. In: Molthen RC, Weaver JB, editors., 2010, p. 762613. https://doi.org/10.1117/12.844587.

[21] Ahmed H, Choy S, Wheatley A, Etemad-Rezai R, Paterson N, Parraga G. Mapping And Quantifying Temporal Dynamics Of Ventilation Heterogeneity In Cystic Fibrosis Using Hyperpolarized Helium-3 Magnetic Resonance Imaging: Developing New Imaging Measurements For Clinical Trials. B70. FINE Anal. Alveolar/airw. Struct. Funct., American Thoracic Society; 2010, p. A3642–A3642. https://doi.org/10.1164/ajrccm-conference.2010.181.1_MeetingAbstracts.A3642.

[22] Bannier E, Cieslar K, Mosbah K, Aubert F, Duboeuf F, Salhi Z, et al. Hyperpolarized 3 He MR for Sensitive Imaging of Ventilation Function and Treatment Efficiency in Young Cystic Fibrosis Patients with Normal Lung Function. Radiology 2010;255:225–32. https://doi.org/10.1148/radiol.09090039.

[23] O’Sullivan B, Couch M, Roche JP, Walvick R, Zheng S, Baker D, et al. Assessment of Repeatability of Hyperpolarized Gas MR Ventilation Functional Imaging in Cystic Fibrosis. Acad Radiol 2014;21:1524–9. https://doi.org/10.1016/j.acra.2014.07.008.

[24] Couch MJ, Thomen R, Kanhere N, Hu R, Ratjen F, Woods J, et al. A two-center analysis of hyperpolarized 129Xe lung MRI in stable pediatric cystic fibrosis: Potential as a biomarker for multi-site trials. J Cyst Fibros 2019;18:728–33. https://doi.org/10.1016/j.jcf.2019.03.005.

[25] Smith LJ, Horsley A, Bray J, Hughes PJC, Biancardi A, Norquay G, et al. The assessment of short and long term changes in lung function in CF using 129

Xe MRI. Eur Respir J 2020:2000441. https://doi.org/10.1183/13993003.00441-2020.

[26] Smith L, Collier G, Marshall H, Hughes P, Biancardi A, Norquay G, et al. P213 A comparison of ventilation MRI using hyperpolarised 3He and 129Xe to assess cystic fibrosis lung disease. J Cyst Fibros 2019;18:S117. https://doi.org/10.1016/S1569-1993(19)30506-5.

[27] Smith L, Collier G, Marshall H, Hughes P, Biancardi A, Norquay G. Comparison of hyperpolarised 3He and 129Xe ventilation MRI to assess lung disease in cystic fibrosis. Int. Soc. Magn. Reson. Med., International Society for Magnetic Resonance in Medicine; 2019.

[28] Koumellis P, van Beek EJR, Woodhouse N, Fichele S, Swift AJ, Paley MNJ, et al. Quantitative analysis of regional airways obstruction using dynamic hyperpolarized3He MRI—Preliminary results in children with cystic fibrosis. J Magn Reson Imaging 2005;22:420–6. https://doi.org/10.1002/jmri.20402.

[29] Mentore K, Froh DK, de Lange EE, Brookeman JR, Paget-Brown AO, Altes TA. Hyperpolarized HHe 3 MRI of the Lung in Cystic Fibrosis. Acad Radiol 2005;12:1423–9. https://doi.org/10.1016/j.acra.2005.07.008.

[30] van Beek EJR, Hill C, Woodhouse N, Fichele S, Fleming S, Howe B, et al. Assessment of lung disease in children with cystic fibrosis using hyperpolarized 3-Helium MRI: comparison with Shwachman score, Chrispin-Norman score and spirometry. Eur Radiol 2007;17:1018–24. https://doi.org/10.1007/s00330-006-0392-1.

[31] Kirby M, Villemaire L, Ahmed H, Paterson NA, McCormack DG, Lewis JF. Diffusion-Weighted Hyperpolarized Helium-3 Magnetic Resonance Imaging In

Adult Cystic Fibrosis. Am J Respir Crit Care Med 2013;194:A2072.

[32] Paulin GA, Svenningsen S, Jobse BN, Mohan S, Kirby M, Lewis JF, et al. Differences in hyperpolarized 3 He ventilation imaging after 4 years in adults with cystic fibrosis. J Magn Reson Imaging 2015;41:1701–7. https://doi.org/10.1002/jmri.24744.

[33] Hardy SM. Study of hyperpolarised 3He MRI diffusion on asthma and cystic fibrosis, and development of hyperpolarised 129Xe MRI lung imaging. University of Nottingham, 2016.

[34] Marshall H, Horsley A, Taylor CJ, Smith L, Hughes D, Horn FC, et al. Detection of early subclinical lung disease in children with cystic fibrosis by lung ventilation imaging with hyperpolarised gas MRI. Thorax 2017;72:760–2. https://doi.org/10.1136/thoraxjnl-2016-208948.

[35] Smith L, Collier G, Marshall H, Hughes P, Biancardi A, Norquay G, et al. P212 Ventilation MRI tracks longitudinal lung function changes in patients with cystic fibrosis and clinically stable FEV1 and Lung Clearance Index. J Cyst Fibros 2019;18:S117. https://doi.org/10.1016/S1569-1993(19)30505-3.

[36] Thomen RP, Walkup LL, Roach DJ, Cleveland ZI, Clancy JP, Woods JC. Hyperpolarized 129Xe for investigation of mild cystic fibrosis lung disease in pediatric patients. J Cyst Fibros 2017;16:275–82. https://doi.org/10.1016/j.jcf.2016.07.008.

[37] Kanhere N, Couch MJ, Kowalik K, Zanette B, Rayment JH, Manson D, et al. Correlation of Lung Clearance Index with Hyperpolarized 129 Xe Magnetic Resonance Imaging in Pediatric Subjects with Cystic Fibrosis. Am J Respir Crit Care Med 2017;196:1073–5. https://doi.org/10.1164/rccm.201611-2228LE.

[38] Rayment JH, Couch MJ, McDonald N, Kanhere N, Manson D, Santyr G, et al. Hyperpolarised 129 Xe magnetic resonance imaging to monitor treatment response in children with cystic fibrosis. Eur Respir J 2019;53:1802188. https://doi.org/10.1183/13993003.02188-2018.

[39] Rayment JH, McDonald N, Kanhere N, Couch M, Santyr G, Ratjen FA. Posture-Dependence of the Lung Clearance Index Varies with Disease State in Cystic Fibrosis. A72. WHAT'S NEW Cyst. FIBROSIS, BPD OTHER Congenit. Pediatr. LUNG Dis., San Diego: American Thoracic Society; 2018, p. A2308–A2308.

[40] Altes T, Johnson MA, Miller GW, Mugler JP, Flors L, Mata J, et al. 46 Hyperpolarized Gas MRI of ivacaftor therapy in subjects with cystic fibrosis who have the G551D-CFTR mutation. J Cyst Fibros 2012;11:S67. https://doi.org/10.1016/S1569-1993(12)60215-X.

[41] Altes TA, Johnson M, Fidler M, Botfield M, Tustison NJ, Leiva-Salinas C, et al. Use of hyperpolarized helium-3 MRI to assess response to ivacaftor treatment in patients with cystic fibrosis. J Cyst Fibros 2017;16:267–74. https://doi.org/10.1016/j.jcf.2016.12.004.

[42] Altes T, Johnson M, Mugler III J, Miller GW, Flors L, Mata J, et al. The Effect Of Ivacaftor, An Investigational CFTR Potentiator, On Hyperpolarized Noble Gas Magnetic Resonance Imaging In Subjects With Cystic Fibrosis Who Have The G551D-CFTR Mutation. B35. Pathog. Clin. ISSUES Cyst. Fibros., American Thoracic Society; 2012, p. A2814–A2814. https://doi.org/10.1164/ajrccm-conference.2012.185.1_MeetingAbstracts.A2814.

[43] Smith LJ, Collier GJ, Marshall H, Hughes PJC, Biancardi AM, Wildman M, et al. Patterns of regional lung physiology in cystic fibrosis using ventilation magnetic resonance imaging and multiple-breath washout. Eur Respir J 2018;52:1800821. https://doi.org/10.1183/13993003.00821-2018.

[44] Smith L, Hughes PJ, Marshall H, Collier G, West N, Wildman M, et al. Hyperpolarised Gas MRI Shows Reversibility of Lung Ventilation Defects at Increased Inspiratory Lung Volumes in Cystic Fibrosis. A72. WHAT'S NEW Cyst. FIBROSIS, BPD OTHER Congenit. Pediatr. LUNG Dis., San Diego: American Thoracic Society; 2018, p. A2313–A2313.

[45] Smith L, Marshall H, Aldag I, Horn F, Collier G, Hughes D, et al. Longitudinal Assessment of Children with Mild Cystic Fibrosis Using Hyperpolarized Gas Lung Magnetic Resonance Imaging and Lung Clearance Index. Am J Respir Crit Care Med 2018;197:397–400. https://doi.org/10.1164/rccm.201705-0894LE.

[46] Smith LJ, Marshall H, Bray J, Wildman M, West N, Horsley A, et al. The effect of acute maximal exercise on the regional distribution of ventilation using ventilation MRI in CF. J Cyst Fibros 2020. https://doi.org/10.1016/j.jcf.2020.08.009.

[47] van Kampen DA, Willems W, van Beers LWAH, Castelein RM, Scholtes VAB, Terwee CB. Determination and comparison of the smallest detectable change (SDC) and the minimal important change (MIC) of four-shoulder patient-reported outcome measures (PROMs). J Orthop Surg Res 2013;8:40. https://doi.org/10.1186/1749-799X-8-40.

[48] McNulty W, Usmani OS. Techniques of assessing small airways dysfunction.

Eur Clin Respir J 2014;1:25898. https://doi.org/10.3402/ecrj.v1.25898.

[49] Francisco B, Ner Z, Ge B, Hewett J, König P. Sensitivity of different spirometric tests for detecting airway obstruction in childhood asthma. J Asthma 2015;52:505–11. https://doi.org/10.3109/02770903.2014.984842.

[50] Hughes PJC, Smith L, Chan H-F, Tahir BA, Norquay G, Collier GJ, et al. Assessment of the influence of lung inflation state on the quantitative parameters derived from hyperpolarized gas lung ventilation MRI in healthy volunteers. J Appl Physiol 2019;126:183–92. https://doi.org/10.1152/japplphysiol.00464.2018.

[51] He M, Driehuys B, Que LG, Huang Y-CT. Using Hyperpolarized 129Xe MRI to Quantify the Pulmonary Ventilation Distribution. Acad Radiol 2016;23:1521–31. https://doi.org/10.1016/j.acra.2016.07.014.

[52] Oren O, Kebebew E, Ioannidis JPA. Curbing Unnecessary and Wasted Diagnostic Imaging. JAMA 2019;321:245. https://doi.org/10.1001/jama.2018.20295.

Appendix

Appendix 1

Table 3: Excluded Studies after Full-text Assessment and Exclusion Reasons

Authors (Date)	No. of multiple publications	Did not have sufficient information about the comparison between CF and healthy individuals.	Did not meet inclusion criteria	Did not assess hyperpolarised gas MRI reliability and validity or no information how it was assessed	Only included one CF patient and there was no sufficient information about the comparison between CF and healthy individuals
Altes at el (2015)	1	*		*	
Carlson et al (2018)	None	*		*	
Donnelly et al (1999)	None		*	*	
Horn et al (2014)	1		*	*	*
Kirby et al (2012)	None		*	*	
Marshall et al (2014)	None	*		*	

Qing et al (2015)	None			*	*
Smith et al (2017)	None		*	*	
Sun et al (2011)	None	*		*	
Thomen et al (2017)	1	*		*	
Tustison et al (2011)	None		*	*	
Youn et al (2012)	None	*		*	
Niedbalski et al (2019)	None		*	*	
Munidasa et al (2019)	None	*		*	

Appendix 2

Table 9: Hypotheses Testing Findings for construct (convergent) validity

Authors (Date)	Study hypothesis	Results	Results in support of hypothesis?

McMahon et al (2006) [4]	“^{3}He MRI would correlate with the major structural abnormalities seen on HRCT and also with functional information provided by spirometry, thus indicating a potential role as a marker of disease status in CF”	^{3}He MRI had strong functional correlation with spirometry and structural CT abnormalities	Yes
Bannier et al (2010) [22]	NA	CF patients had ventilation defects, even though spirometry results showed normal lung function.	NA
Kirby et al (2011) [18]	“He MRI would provide the necessary and sufficient spatial and temporal sensitivity to detect day-to-day changes in lung function”	The results showed changes in ventilation defects when the ^{3}He-MRI repeated after 7 days, but day to day changes in the lung was not assessed.	No
Altes et al (2012) [40]	“^{3}He-MRI would be appropriate for evaluating response to ivacaftor”	^{3}He MRI was able to detect the lungs response to invacaftor which was effective in improving lung ventilation in CF patients	Yes
Smith et al (2018) [43]	NA	VDP strongly correlated with lung clearance index and forced expiratory volume in 1 s (FEV1)	NA
Zha et al (2019) [8]	“Oxygen enhanced MRI may yield comparable whole-lung VDP relative to hyperpolarized ^{3}He MRI as the reference method”	OE MRI showed similar performance compared with ^{3}He MRI for measuring VDP	Yes

Smith et al (2018) [45]	NA	Ventilation MRI is capable of detecting significant lung function changes in the follow-up of children with CF and normal spirometry	NA
Smith et al, (2019) [26]	NA	There was no inherent bias for VDP between the two gases although at an individual level differences were evident. Despite this, when followed up both gases similarly reflected changes in ventilation, suggesting both are capable of reflecting CF lung disease severity.	NA