Editorial Commentary

Transbronchial cryobiopsy in the diagnosis of interstitial lung diseases: methodologies and perspectives from the Cryo-PID and COLDICE studies

Carey Suehs[1], Arnaud Bourdin[1,2], Isabelle Vachier[1], Nicolas Molinari[3], Micaela Romagnoli[4]

[1]Department of Respiratory Diseases, Univ Montpellier, CHU Montpellier, Montpellier, France; [2]PhyMedExp, Univ Montpellier, CNRS, INSERM, CHU Montpellier, Montpellier, France; [3]IMAG, CNRS, Univ Montpellier, CHU Montpellier, Montpellier, France; [4]Pulmonology Unit, Ospedale Ca' Foncello, AULSS2 Marca Trevigiana, Treviso, Italy

Correspondence to: Carey Suehs. Department of Respiratory Diseases, Univ Montpellier, CHU Montpellier, Hôpital Arnaud de Villeneuve, 371 Av. du Doyen Gaston Giraud, F-34090 Montpellier, France. Email: careysuehs@protonmail.com.

Comment on: Troy LK, Grainge C, Corte TJ, *et al.* Diagnostic accuracy of transbronchial lung cryobiopsy for interstitial lung disease diagnosis (COLDICE): a prospective, comparative study. Lancet Respir Med 2020;8:171-81.

Submitted Mar 30, 2020. Accepted for publication Apr 27, 2020.
doi: 10.21037/atm-20-2814
View this article at: http://dx.doi.org/10.21037/atm-20-2814

For a substantial majority of diffuse interstitial lung disease (ILD) cases, a multidisciplinary discussion (MDD) of clinical and imaging data results in a diagnosis. For 30–40% of these patients, however, a first MDD (MDD1) indicates that further data are required, i.e., histopathological interpretation of a lung biopsy (1-4). This is particularly true when no clear "working diagnosis" is proposed, most frequently because of "unclassifiable" high resolution computed tomography (HRCT) patterns. A dilemma then ensues: do the potential health-benefits resulting from an improvement in diagnostic accuracy justify the health-risks associated with the biopsy procedure? Indeed, the stakes for the patient are high, with certain histopathological parenchyma patterns that are visually quite similar and often confused corresponding to radically different prognoses and eventually treatments [typically idiopathic pulmonary fibrosis (IPF) versus non-IPF]. Additionally, the ILD patient population is often comorbid and may or may not be able to tolerate surgical procedures. In borderline patients, the means of increasing the safety profile of lung biopsy procedures takes on particular interest, given that the diagnostic accuracy of the resulting tissue samples stays the same.

Within this rather complex framework, two recently published studies both aimed to evaluate the diagnostic accuracy of transbronchial lung cryobiopsy (TBLC) as a potentially safer (5,6) means of obtaining parenchyma samples as compared to the gold standard of surgical lung biopsy (SLB). The first-published (the "Cryo-PID" study) was a bicentric European rather small-sample-size study (N=21) that concluded in poor diagnostic concordance between TBLC and SLB samples (7). The second ("COLDICE" study) was a larger multicenter Australian study (N=65) that came to the opposite conclusion of good diagnostic concordance between the two biopsy procedures (8). These two opposite conclusions should suggest to the readers at least two main thoughts: first of all, results from these two studies are controversial despite the use of good methodologies in both studies, and secondly sample size but especially statistical analyses might have influenced such a difference in the final message. The reader should also immediately note that the authors of the present invited editorial commentary are members of the team that produced the small-sample-size study (7). Our aim is to discuss the results of either study and how they may be transposed to real-life clinical contexts, with multiple perspectives on how to move forward in the domain. Of note, TBLC are already routinely performed in different settings/centers, suggesting a real infatuation of bronchoscopists for this innovation. Given the substantial size of the COLDICE study and its strong, positive results, it is worth closely analyszing how they were generated and what interpretations may be made.

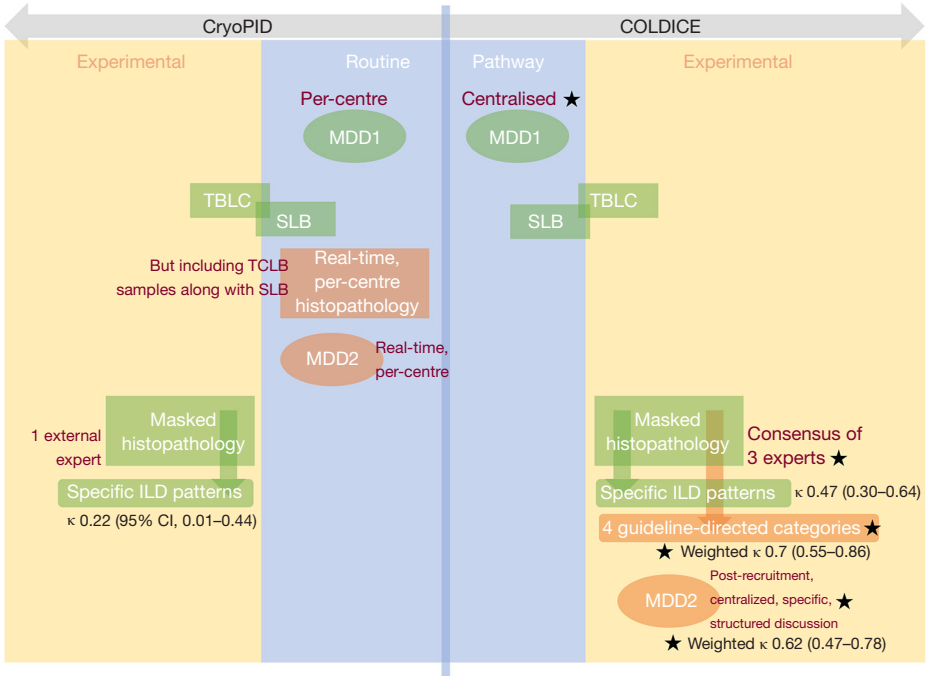

Figure 1 Graphic presentation of the methodological similarities (indicated in green) and differences (indicated in red) for two recently published studies (Cryo-PID) (7) and COLDICE (8,9) evaluating the diagnostic accuracy of TBLC against that of the gold standard (SLB) for ILD diagnosis. Both studies shared the same general goal and structure, with (I) a MDD1 followed by (II) the sequential performance of TBLC and SLB at the same anatomical locations and in the same surgical session for a given patient, (III) subsequent masked histopathology to compare the two biopsy types in a blinded manner and (IV) a second MDD2. However, MDDs were performed within the framework of per-center routine practice in the Cryo-PID study, whereas the COLDICE study used MDDs centralized over their nine participating centers (with MDD2 occurring after the study recruitment period and with a research-specific structured discussion). Masked histopathology in Cryo-PID was performed by internationally recognized external expert, while that of COLDICE reported the consensus of 3 experts. Both of the latter histological evaluations produced specific ILD patterns, but only COLDICE further provided guideline-directed categories. TBLC versus SLB concordance estimates (κ) are given in black. Stars (★) indicate elements that can increase estimates of concordance either by reducing population or evaluation heterogeneity (centralization of MDDs over nine centers, consensus estimates for histology), or automatically through probability and math (using a smaller number of categories or using weighted κ calculations). TBLC, transbronchial lung cryobiopsy; SLB, surgical lung biopsy; ILD, interstitial lung disease; MDD1, first multidisciplinary discussion.

Figure 1 summarizes the methodological similarities and differences between Cryo-PID and COLDICE. First, these two studies are one-of-a-kind in being the first to perform sequential TBLC and SLB at the same anatomical sites within a single surgical session for a given patient. However, subsequent methodological differences render the biopsy concordance results reported by the two studies incomparable. This unfortunate situation however informs on the range of TBLC-*vs*.-SLB concordance values that can be expected in real life clinical situations, which provides food-for-thought. It also highlights how different methodological choices can affect concordance results, and why this is important.

Non-weighted (κ) versus weighted (κw) estimates of concordance

In general, weighted concordance estimates (κw) give higher numbers than their non-weighted counterparts (κ), and the two are thus incomparable. This occurs for two reasons: (I) the ranking schemes used have fewer classes than specific disease patterns and (II) the math involved takes into account the notion of rank. A good example of the first case is the system used by Thomeer *et al*. (10), which ranks histological patterns as “unlikely”, “probable”, or “very suggestive”. A second example is the ranks recommended by the relevant 2018 ATS/ERS/JRS/ALAT

Clinical Practice Guideline (3) and used in the COLDICE study: "definite usual interstitial pneumonia", "probable usual interstitial pneumonia", "indeterminate for usual interstitial pneumonia", and "alternative diagnosis". These ranking schemes have fewer classes than the classified histo-pathological patterns by Travis *et al*. (11) assessed and used in the Cryo-PID study (e.g., UIP, NSIP, HP, DIP, LIP, PPFE, etc.), thus increasing the probability that two different experts will assign the same result to a given biopsy. One can further expect that the fewer the number of classes in a ranking scheme, the more likely different experts will give the same result. Additionally, because weighted kappa estimates take into account a quantitative aspect among classes, i.e., a "rank"—meaning that certain categories are more distant than others—κw estimates are typically higher than their non-weighted κ counterparts. In addition, how rankings are created (e.g., via linear or quadratic methods) can affect results and should be discussed/defined in a priori fashion. The higher values portrayed by κw are not incorrect, but they are used differently, and the reader should keep in mind that they are not comparable to simple κ. Attempting to compare a κ with a κw would, at best, be misleading.

A good example of the difference between κ and κw for a given dataset is provided in the COLDICE study, who found a κ of 0.46 and a κw of 0.7 when comparing the histopathological results of TBLC and SLB, and yet both estimates are accompanied by extremely similar percentages of agreement (A69.2% *vs*. A70.8%, respectively). The increase in the concordance estimate relative to the % agreement is simply the result of a classification with a smaller number of categories that are furthermore assigned semi-quantitative ranks. If we class quintiles of concordance estimates as "poor" [κ(w) ≤0.2], "fair" [0.2< κ(w) ≤0.4], "moderate" [0.4< κ(w) ≤0.6], "good" (0.6< κ(w) ≤0.8) and "excellent" concordance [0.80< κ(w) ≤1.00], as in previous studies (4,8,12), the COLDICE estimate for concordance between TBLC and SLB jumps from "moderate" to "good" simply depending on the calculation used.

Inter-observer agreement

In this notoriously difficult domain, the concordance between two different experts for a first-choice diagnosis has been previously demonstrated as only "fair" [κ 0.31 (4); κ 0.38 (12)]. This is a major source of heterogeneity that must be taken into account when comparing the histopathological scoring of TBLC versus SLB tissue samples. These estimates should be kept in mind because they indicate how two different experts would classify the same biopsy. One should hope, logically, that these estimates, for a single biopsy, are larger than those when comparing two different biopsies (i.e., inter-sample concordance or agreement). They represent therefore a logical upper-bound in what can be expected for TBLC versus SLB simple concordance estimates (κ).

Unfortunately, the COLDICE study results for inter-observer concordance cannot be compared to the latter estimates. Indeed, they provide again yet a different type of concordance estimate, i.e., the Fleiss' kappa (κ_F), which is used when there are more than two experts involved. Average two-by-two expert concordances (κ) or % agreement that could be compared with previous studies are not provided (8). Curiously, when using the ATS/ERS/JRS/ALAT 4-category ranking system in COLDICE, the κ_F for TBLC barely increased from κ_F 0.52 to κ_F 0.53, while that for SLB increased from κ_F 0.50 to κ_F 0.64. This suggests that when histopathology experts gave different results at the individual tissue pattern level, these differences often span the ATS/ERS/JRS/ALAT categories in TBLCs and less so in SLBs.

Pathology results based on a single expert, or a consensus from multiple experts?

A major methodological difference between Cryo-PID and COLDICE is that Cryo-PID used a single, external, expert pathologist for evaluating the concordance between TBLC and SLB samples from a given patient, while COLDICE uses the consensus from 3 experts (*Figure 1*). This difference is likely to reduce the heterogeneity innate to experts discussed in the previous section (just like regression towards the mean) and helps explain the increase in agreement estimates in COLDICE (κ 0.47, A69.2%) relative to Cryo-PID (κ 0.22, A38%). From the reader's point-of-view, the most relevant concordance estimate depends on which situation most-resembles his/her routine practice. In France and Italy, where the Cryo-PID study took place, histopathology is performed in routine by a single expert. It is unlikely in this context that TBLC-SLB simple concordance (κ) estimates in our area will ever reach the 3-expert-consensus based κ or κw levels reported in COLDICE.

Single center or centralized MDDs?

Agreement between different MDD teams for first-choice

diagnosis has been described as only moderate (κ 0.5) (4). As for inter-observer agreement, this represents a source of heterogeneity, and the choice of centralized MDDs used in COLDICE again helps to explain the high agreement reported in their study. MDDs were centralized over the nine participating centers in COLDICE, avoiding a considerable source of heterogeneity. Furthermore, the COLDICE MDD2 was organized as a single-session, post-recruitment mass review of cases. Though steps were taken to avoid case-memorization bias in this situation, the latter cannot be ruled out and does not correspond to routine practice.

Agreement in real life

Overall, several methodological choices in COLDICE help explain the purportedly high agreement between TBLC and SLB samples they reported (*Figure 1*). By reducing potential sources of heterogeneity, they optimized the potential of pathology experts and MDDs for giving accurate results. This is a useful property as the "moderate" κ 0.47 for TBLC-SLB agreement reported in COLDICE likely therefore marks a high-end estimate of κ. Troy *et al.* themselves state "*It is uncertain if TBLC accuracy will be similar in the wider clinical setting*" (8), implying that real-life TBLC-SLB agreement is likely to be lower. In contrast, the small sample size of Cryo-PID is a serious handicap; the associated κ 0.22 may be a low-estimate simply due to chance. Most likely, this estimate is however more representative of centers with local MDA and expert histopathologists, as is common in routine practice.

Nevertheless, COLDICE should be lauded as the only current estimate of TBLC-SLB histopathological agreement using the ATS/ERS/JRS/ALAT Clinical Practice Guideline (3) recommended categories. The latter provides a clinically meaningful, simplified ranking system for lung biopsies. However, the associated, seemingly high, κw of 0.7, the primary result of the COLDICE study, is specific to this ranking system and likely a high-end estimate of κw due to the research-specific (non-routine) procedures used.

The ideal diagnostic accuracy study

In hindsight, the ideal diagnostic accuracy study establishing agreement statistics between TBLC and SLB would use both the Cryo-PID and the COLDICE methodologies (elements in *Figure 1*). By juxtaposing per-center MDD results with centralized ones, single-expert histopathology results with consensuses, and simple κ statistics with weighted κw ones, not only would diagnostic accuracy be established on several levels adaptable to several scenarios, but the variation due to different sources of heterogeneity could be established. The associated gain via multi-expert and teamwork environments could be quantified, perhaps justifying their deployment to routine practice.

Perspectives and conclusion

Future studies in the domain may also include the development of scientific consensus for appropriate outcomes and methodology, and consideration of interventional studies randomizing TBLC against SLB for homogeneity of results and safety. Considering the latter, bleeding events in TBLC have been reported as a factor contributing to higher mortality in ILD patients (13). The soiling of the airways by blood can reduce lung function parameters and appropriate measures should be implemented for bleeding prophylaxis (13). This is not a care-free matter if TBLC is to be developed for borderline situations where SLB risk/benefit ratios were estimated to be negative.

In conclusion, real-life concordance (κ) between TBLC and SLB is likely to fall between the two estimates provided by the two main studies conducted in the field. However, the real challenges are likely to be elsewhere. First, will TBLC be feasible and provide additional benefits to patients who are ineligible for SLB? Second, what are the mechanistic consequences for research, which will suffer from smaller biopsy sizes and increased diagnostic uncertainties? And last but not least, what is the future of these very debates in the world of biologics, where hopefully disease-specific molecular targets and corresponding immunotherapies will be identified?

Acknowledgments

Funding: None.

Footnote

Provenance and Peer Review: This article was commissioned by the editorial office, *Annals of Translational Medicine.* The article did not undergo external peer review.

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at http://dx.doi.org/10.21037/atm-20-2814). CS reports grants from

Astra Zeneca, outside the submitted work. AB reports grants, personal fees, non-financial support and other from AstraZeneca, grants, personal fees, non-financial support and other from Boeringher Ingelheim, grants, personal fees, non-financial support and other from GlaxoSmithKline, personal fees, non-financial support and other from Novartis, personal fees and non-financial support from Teva, personal fees, non-financial support and other from Regeneron, personal fees, non-financial support and other from Chiesi Farmaceuticals, grants, personal fees, non-financial support and other from Actelion, personal fees from Gilead, non-financial support and other from Roche, other from Nuvaira, outside the submitted work. NM reports grants and personal fees from AstraZeneca, outside the submitted work. MR reports grants and personal fees from AstraZeneca, grants and personal fees from GSK, grants from Boeringher, grants from Menarini, grants and personal fees from Roche, grants from Chiesi, grants from Alfasigma, grants and personal fees from Novartis, personal fees from Guidotti, outside the submitted work. IV has no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

References

1. King TE Jr, Bradford WZ, Castro-Bernardini S, et al. A phase 3 trial of pirfenidone in patients with idiopathic pulmonary fibrosis. N Engl J Med 2014;370:2083-92.
2. Richeldi L, du Bois RM, Raghu G, et al. Efficacy and safety of nintedanib in idiopathic pulmonary fibrosis. N Engl J Med 2014;370:2071-82.
3. Raghu G, Collard HR, Egan JJ, et al. An official ATS/ERS/JRS/ALAT statement: idiopathic pulmonary fibrosis: evidence-based guidelines for diagnosis and management. Am J Respir Crit Care Med 2011;183:788-824.
4. Walsh SLF, Wells AU, Desai SR, et al. Multicentre evaluation of multidisciplinary team meeting agreement on diagnosis in diffuse parenchymal lung disease: a case-cohort study. Lancet Respir Med 2016;4:557-65.
5. Lentz RJ, Argento AC, Colby TV, et al. Transbronchial cryobiopsy for diffuse parenchymal lung disease: a state-of-the-art review of procedural techniques, current evidence, and future challenges. J Thorac Dis 2017;9:2186-203.
6. Maldonado F, Danoff SK, Wells AU, et al. Transbronchial cryobiopsy for the diagnosis of interstitial lung diseases: CHEST guideline and expert panel report. Chest 2020;157:1030-42.
7. Romagnoli M, Colby TV, Berthet JP, et al. Poor concordance between sequential transbronchial lung cryobiopsy and surgical lung biopsy in the diagnosis of diffuse interstitial lung diseases. Am J Respir Crit Care Med 2019;199:1249-56.
8. Troy LK, Grainge C, Corte TJ, et al. Diagnostic accuracy of transbronchial lung cryobiopsy for interstitial lung disease diagnosis (COLDICE): a prospective, comparative study. Lancet Respir Med 2020;8:171-81.
9. Troy LK, Grainge C, Corte T, et al. Cryobiopsy versus open lung biopsy in the diagnosis of interstitial lung disease (COLDICE): protocol of a multicentre study. BMJ Open Respir Res 2019;6:e000443.
10. Thomeer M, Demedts M, Behr J, et al. Multidisciplinary interobserver agreement in the diagnosis of idiopathic pulmonary fibrosis. Eur Respir J 2008;31:585-91.
11. Travis WD, Costabel U, Hansell DM, et al. An official American Thoracic Society/European Respiratory Society statement: update of the international multidisciplinary classification of the idiopathic interstitial pneumonias. Am J Respir Crit Care Med 2013;188:733-48.
12. Nicholson AG, Addis BJ, Bharucha H, et al. Inter-observer variation between pathologists in diffuse parenchymal lung disease. Thorax 2004;59:500-5.
13. Pannu J, Roller LJ, Maldonado F, et al. Transbronchial cryobiopsy for diffuse parenchymal lung disease: 30- and 90-day mortality. Eur Respir J 2019. doi: 10.1183/13993003.00337-2019.

Cite this article as: Suehs C, Bourdin A, Vachier I, Molinari N, Romagnoli M. Transbronchial cryobiopsy in the diagnosis of interstitial lung diseases: methodologies and perspectives from the Cryo-PID and COLDICE studies. Ann Transl Med 2020;8(20):1330. doi: 10.21037/atm-20-2814

