ANNALS OF INTERNAL MEDICINE - DATA SHARING AND RE-USE
STATISTICAL ANALYSIS PLAN

The two databases have been completed as laid down in our preregistered protocol (https://osf.io/gnt6u/). Both of them have been checked by independent reviewers (final check: 07/15/2019).
Briefly, the first database gathers data about the Annals of Internal Medicine's randomized controlled trials (RCT) from 2007 to 2017: article's PMID, title, authors' names, date of publication and corresponding author's geographical location; trial's intervention type, type of control group, medical specialty, sample size (achieved at the end of trial recruitment), result on the primary outcome, funding source, and sharing status of the dataset, statistical code and study protocol; number of citations from Web of Science.
The second one gathers data about the reuses of those RCT: reuse's PMID, title, authors, date of publication and type (re-analysis, secondary analysis, meta-analysis on individual patient data (IPD) or pooled analysis); mention about the availability of the reused data; although not required in our protocol, the meta-analysis on aggregate data also have been included in the database. In addition, we clarified on point from our initial protocol: we excluded citations concerning long term follow up and use of stored biological samples since these analyses obviously implied acquisition of new data.

For quality and transparency purposes, we chose our models before merging the two databases and pre-specified these choices in the following statistical analysis plan. Only the first descriptive elements (cf. Tables 1, 2, 3, 4 and Figure 1) were analyzed before writing of this document.

All statistical tests will be performed with a 2-sided hypothesis at 5% alpha level.

Table 1. Characteristics of the included RCT

	Overall	Unshared dataset	Shared dataset	p
Number of articles	185	79	106	
Year of publication median [IQR]	2012 [2010, 2015]	2012 [2010, 2014.5]	2012 [2009, 2015]	0.797*
Geographical zones n(%)				0.318†
Asia	13 (7.0)	8 (10.1)	5 (4.7)	
Europe	54 (29.2)	20 (25.3)	34 (32.1)	
Other	21 (11.4)	7 (8.9)	14 (13.2)	
USA	97 (52.4)	44 (55.7)	53 (50.0)	
Intervention type n(%)				0.620†
Complex intervention	89 (48.1)	35 (44.3)	54 (50.9)	
Device	8 (4.3)	3 (3.8)	5 (4.7)	
Drug	88 (47.6)	41 (51.9)	47 (44.3)	
Control group n(%) **Ref: inactive**	140 (75.7)	61 (77.2)	79 (74.5)	0.804
Medical specialty n(%) **Ref: surgery**	4 (2.2)	3 (3.8)	1 (0.9)	0.314†
Primary outcome n(%) **Ref: positive**	123 (66.5)	58 (73.4)	65 (61.3)	0.117
Funding source n(%)				0.200†
Academic	79 (42.7)	31 (39.2)	48 (45.3)	
Charity	10 (5.4)	3 (3.8)	7 (6.6)	
Industry	28 (15.1)	17 (21.5)	11 (10.4)	
Mixed	68 (36.8)	28 (35.4)	40 (37.7)	
Sample size median [IQR]	299 [158, 719]	267.00 [151, 663]	310 [170.25, 722.75]	0.442*
Protocol shared n(%) **Ref: yes**	151 (81.6)	51 (64.6)	100 (94.3)	<0.001
Code shared n(%) **Ref: yes**	134 (72.4)	43 (54.4)	91 (85.8)	<0.001
Number of citations median [IQR]	59 [29, 114]	71 [36, 116.5]	51.5 [25, 112.75]	0.082*

* : Kruskal-Wallis test

† : Fisher exact test

Table 2. Characteristics of reuses of the 185 included RCT

Reuses taken into account:

- re-analysis, secondary analysis, meta-analysis on IPD and pooled analysis

- without or with the primary outcome's authorship restriction

		Never reused	Reused at least once
Number of RCT n(%)	**Without authorship restriction**	94 (51 %)	91 (49 %)
	With authorship restriction	Still unknown (databases not merged)	Still unknown (databases not merged)

Table 3. Number and types of reuses of the 185 included RCT

Reuses taken into account:

- re-analysis, secondary analysis, meta-analysis on IPD and pooled analysis

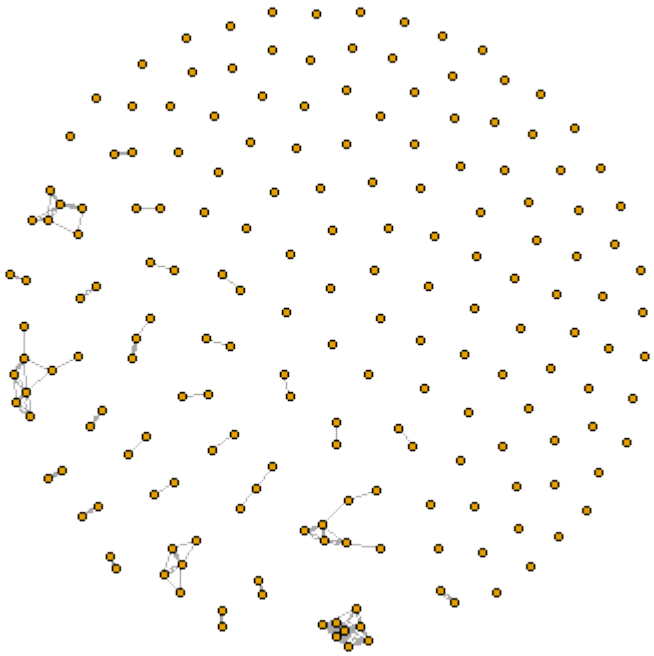
- without the primary outcome's authorship restriction

		Number of RCT which are reused 0, 1, 2, 3 or more times				
		0 reuse	**1 reuse**	**2 reuses**	**3 reuses**	**≥ 4 reuses**
Type of reuse	**Re-analysis** n (%)	0	0	0	0	0
	Secondary analysis n (%)	102 (55)	31 (17)	23 (12)	9 (5)	20 (11)
	Meta-analysis on individual patient data or pooled analysis n (%)	161 (87)	13 (7)	5 (3)	1 (0.5)	5 (2.5)
	Any type n (%)	94 (51)	29 (16)	27 (15)	8 (4)	27 (15)

Figure 1. Network identifying multiple RCT's publications from the same team (at least one author in common)

1 node = 1 article

1 edge = 1 author in common between the linked articles



I. DATA SHARING RATES OVER TIME

We will first perform an explanatory analysis describing the intent to share data rates over time. As planned in the protocol, this association will be adjusted on several factors.

General strategy for modeling	
Dependent variable	Intent to share (binary variable: yes/no)
Explanatory variable of interest	Time (continuous variable, or qualitative if required)
Other explanatory variables	Geographical zone, intervention type, control group type, medical specialty, result on the primary outcome, funding source, sample size
Model	Logistic regression
Association measure	Odds ratio and its 95% confidence interval
Method	1) A univariate model will be used to explore the relationship between the dependent and explanatory variables 2) Effect of the explanatory variable of interest and of all associations with a p-value <0.25 will be subsequently explored in a multivariate model 3) We will perform a mixed model with a random effect for team (primary publications of RCT with at least one author in common) and assess the relevance of this approach by comparing the random effect variance with the residual variance.
Hypothesis and model validation	Log-linearity of the time will be explored graphically; this variable will be transformed or categorized if required. Multicollinearity (variance inflation factor) Testing influential observations (Dfbeta values) • If observations are very influential, the dataset will be checked for false values; these will be corrected if required • No outlier will be suppressed

II. MAIN OUTCOMES

Protocol section	Outcome type	Outcome description	Model
3	Primary	Reuse of the data (limited to citing papers where the first, last and corresponding authors are not among the authors of the original paper)	Survival
4.1	Secondary	Components (1, 2, and 3) of the primary outcome	Survival
4.2	Secondary	Reuse of the data without the authorship restriction used for the primary outcome	Survival
4.3 **4.4** **4.5** **4.6** **4.7**	Secondary	Number and type of secondary use - in which lead authors (first, corresponding and last) of the paper were outside the team authoring the original Annals paper - by a team where 0, <50%, ≥50% or 100% of authors are from the original Annals paper	Count outcome
4.8	Secondary	Mention in the secondary papers about the availability of the reused data and how	Only descriptive data
4.9	Secondary	Number of citations	Count outcome

Survival models

3	Primary	Reuse of the data (limited to citing papers where the first, last and corresponding authors are not among the authors of the original paper)
Dependent variable		Data reused in any secondary use (censored variable / from date of publication to date of citation searches)
Explanatory variable of interest		Intent to share (binary variable)
Other explanatory variables		Geographical zone, intervention type, control group type, medical specialty, result on the primary outcome, funding source, sample size

4.1	Secondary	Components (1, 2, and 3) of the primary outcome
Dependent variable		2 models will be performed: 1/ Data reused in a secondary analysis (censored variable / from date of publication to date of citation searches) 2/ Data reused in a meta-analysis on individual patient data or pooled analysis (censored variable / from date of publication to date of citation searches) As there was no re-analysis, this outcome won't be explored by modelling.
Explanatory variable of interest		Intent to share (binary variable)
Other explanatory variables		Geographical zone, intervention type, control group type, medical specialty, result on the primary outcome, funding source, sample size

4.2	Secondary	Reuse of the data without the authorship restriction used for the primary outcome
Dependent variable		Data reused in any secondary use (censored variable / from date of publication to date of citation searches)
Explanatory variable of interest		Intent to share (binary variable)
Other explanatory variables		Geographical zone, intervention type, control group type, medical specialty, result on the primary outcome, funding source, sample size

General strategy for modeling	
Model	Cox proportional hazards regression model
Association measure	Hazard ratio and its 95% confidence interval
Method	1) A univariate model will be used to explore the relationship between the dependent and explanatory variables 2) Effect of the explanatory variable of interest and of all associations with a p-value <0.25 will be subsequently explored in a multivariate model 3) We will perform a mixed model with a random effect for team (primary publications of RCT with at least one author in common) and assess the relevance of this approach by comparing the random effect variance with the residual variance.
Hypothesis and model validation	Proportional hazards assumption (Schoenfeld residuals, log(-log(S(t))) versus time plots) • If not respected: we will add a variable*time interaction, or stratify or remove the variable Testing influential observations (Dfbeta values) • If observations are very influential, the dataset will be checked for false values; these will be corrected if required • No outlier will be suppressed Linearity (Deviance residual) • If not respected: we will categorize the variable

Count outcome models

4.3 **4.4** **4.5** **4.6** **4.7**	Secondary	Number and type of secondary use - in which lead authors (first, corresponding and last) of the paper were outside the team authoring the original Annals paper - by a team where 0, <50%, ≥50% or 100% of authors are from the original Annals paper
Dependent variable		2 models will be performed for each author criteria: 1/ Number of reuses in a secondary analysis 2/ Number of reuses in a meta-analysis on individual patient data or pooled analysis As there was no re-analysis, this outcome won't be explored by modelling.
Explanatory variable of interest		Intent to share (binary variable)
Other explanatory variables		Year of publication (will be taken into account using an offset corresponding to the difference between publication date and date of citation searches), geographical zone, intervention type, control group type, medical specialty, result on the primary outcome, funding source, sample size

4.9	Secondary	Number of citations
Dependent variable		Number of citations
Explanatory variable of interest		Intent to share (binary variable)
Other explanatory variables		Year of publication (will be taken into account using an offset corresponding to the difference between publication date and date of citation searches), geographical zone, intervention type, control group type, medical specialty, result on the primary outcome, funding source, sample size

General strategy for modeling	
Model	1st intention: Poisson model 2nd intention: in case of overdispersion, a negative binomial regression
Association measure	Incidence rate ratio (IRR) and its 95% confidence interval
Method	1) A univariate model will be used to explore the relationship between the dependent and explanatory variables 2) Effect of the explanatory variable of interest and of all associations with a p-value <0.25 will be subsequently explored in a multivariate model 3) The dispersion will be tested; in case of overdispersion, the model selection will continue from step 1 with a negative binomial regression 4) We will perform a mixed model with a random effect for team (primary publications of RCT with at least one author in common) and assess the relevance of this approach by comparing the random effect variance with the residual variance.
Hypothesis and model validation	No overdispersion Linearity: log of the mean rate must be a linear function of x

III. CHANGES TO THE INITIAL PROTOCOL PRE-SPECIFIED ANALYSIS

Pooled analysis

- The pooled analysis will be considered as meta-analysis on IPD due to their close resemblance.

Year of publication for main outcomes

- In survival models, the RCT's year of publication will be used to calculate each follow-up time (date of publication to date of citation searches).
- In count outcome models, the RCT's year of publication will be taken into account using an offset (corresponding to the difference between publication date and date of citation searches).

Sample size of the RCT

- Since the sample size could be responsible for changes in both the intent to share and the number of reuses, it could be an important confounding factor. In order to deal with it, the sample size will be added in the "other explanatory variables" for all analysis.

Meta-analysis on aggregate data

- We will add a specific analysis for meta-analysis on aggregate data, since data availability could also have an impact when the publication does not report the required data to perform the meta-analysis.
- We will apply the same methodology as for the pre-specified analysis.
    - **Section 3**: Reuse of the data (limited to citing papers where the first, last and corresponding authors are not among the authors of the original paper)
        - We will add the meta-analysis to the reuses that were already included (re-analysis, secondary analysis, meta-analysis on IPD or pooled analysis).
    - **Section 4.1**: Components (1, 2, and 3) of the primary outcome
        - A third model will be performed with the data reused in a meta-analysis on aggregate data as the dependent variable.
    - **Section 4.2**: Reuse of the data without the authorship restriction used for the primary outcome
        - We will add the meta-analysis to the reuses that were already included (re-analysis, secondary analysis, meta-analysis on IPD or pooled analysis).
    - **Sections 4.3 to 4.7**: Number and type of secondary use
        - Two more models for each author criteria will be performed with the data reused in a meta-analysis on aggregate data as the dependent variable

Table 4. Number of meta-analysis on aggregate data of the 185 included RCT

Reuses taken into account:

- meta-analysis on aggregate data

- without the primary outcome's authorship restriction

		Number of RCT which are reused 0, 1, 2, 3 or more times				
		0 reuse	**1 reuse**	**2 reuses**	**3 reuses**	**≥ 4 reuses**
Type of reuse	**Meta-analysis on aggregate data** n (%)	63 (34)	40 (22)	31 (17)	19 (10)	32 (17)

IV. ACKNOWLEDGEMENT

A study research grant was provided from the FRM (Fondation pour la Recherche Médicale). The sponsor had no role concerning preparation, review, or approval of the analysis plan.

We thank Bruno Falissard who provided strong support in writing this statistical analysis plan.