

1 Clinical data specification and coding for cross-analyses with omics data in
2 autoimmune disease trials.

3

4 Lorenzon Roberta^{1,2*}, Drakos Iannis^{1,2*}, Claire Ribet², Sophie Harris², Cordoba Maeva², Tran
5 Olivia³, Dasque Eric⁴, Cacoub Patrice^{5,1,2}, Hartemann Agnes^{6,2}, Bodaghi Bahram^{7,1}, Saadoun
6 David^{5,1,2}, Berenbaum Francis^{8,2}, Grateau Gilles^{9,2}, Ronco Pierre^{10,2}, Benveniste Olivier^{11,2},
7 Mariampillai Kuberaka^{11,2}, Sellam Jeremie^{8,2}, Seksik Philippe^{12,2}, Rosenzweig Michelle^{1,2}, Six
8 Adrien^{1,2}, Bernard Claude², Aheng Caroline², Vicaut Eric³, Klatzmann David^{1,2,**}, Mariotti-
9 Ferrandiz Encarnita^{1,2**}

10 *,** these authors equally contributed to the work

11 ¹ Sorbonne Université, INSERM, UMR S 959, Immunology-Immunopathology-
12 Immunotherapy (I3); F-75005, Paris, France;

13 ² Biotherapy (CIC-BTi) and Inflammation-Immunopathology-Biotherapy Department (DHU
14 i2B), Hôpital Pitié-Salpêtrière, AP-HP, F-75651, Paris, France;

15 ³ Unité de recherche clinique, UMR 942, Univ Paris 07, Hôpitaux Saint Louis Lariboisière,
16 APHP, Paris, France;

17 ⁴ CIC-1421, Hôpital Pitié-Salpêtrière, AP-HP, Paris, France;

18 ⁵ UMR 974, UPMC, Department of Internal Medicine and Clinical Immunology, Hôpital Pitié-
19 Salpêtrière, AP-HP, Paris, France;

20 ⁶ Department of Diabetology, Hôpital Pitié-Salpêtrière, AP-HP, France; Faculty of Medicine,
21 Sorbonne Université; Institute of Cardiometabolism and Nutrition (ICAN), Paris, France;

22 ⁷ Département Hospitalo-Universitaire Vision and Handicaps 'ViewMaintain', Pitié-
23 Salpêtrière University Hospital, Paris, France

24 ⁸ Department of Rheumatology, Sorbonne Université, INSERM UMR_S938, Hôpital Saint-
25 Antoine, APHP, Paris, France;

26 ⁹ Sorbonne Université, INSERM, UMR_S 933, Department of Internal Medicine Hôpital
27 Tenon, APHP, Paris, France;

28 ¹⁰ Sorbonne Université; INSERM, UMR_S 1155, Paris, France; Hôpital Tenon, AP-HP,
29 Department of Nephrology and Dialysis, Paris, France;

30 ¹¹UMR 974, UPMC, Department of Internal Medicine and Clinical Immunology, Hôpital Pitié-
31 Salpêtrière, AP-HP, Paris, France;

32 ¹²Department of Gastroenterology, Hôpital Saint Antoine, AP-HP, and GRC-UPMC 03,
33 Sorbonne Université, Paris, France;

34

35 Keywords: multi-disease; harmonization; clinical trial; clinical coding

36

37 **ABSTRACT**

38 **Objectives:** Autoimmune and inflammatory diseases (AIDs) form a continuum of
39 autoimmune and inflammatory diseases, yet AIDs' nosology is based on syndromic
40 classification. The TRANSIMMUNOM trial (NCT02466217) was designed to re-evaluate AIDs
41 nosology through clinic-biological and multi-omics investigations of patients with one of 19
42 selected AIDs. To allow cross-analyses of clinic-biological data together with omics data, we
43 needed to integrate clinical data in a harmonized database. **Materials and Methods:** We
44 assembled a clinical expert consortium (CEC) to select relevant clinic-biological features to
45 be collected for all patients and a cohort management team comprising biologists, clinicians
46 and computer scientists to design an electronic case report form (eCRF). The eCRF design
47 and implementation has been done on OpenClinica, an open-source CFR-part 11 compliant
48 electronic data capture system. **Results:** The CEC selected 865 clinical and biological
49 parameters. The CMT selected coded the items using CDISC standards into 5835 coded
50 values organized in 28 structured eCRFs. Examples of such coding are check boxes for
51 clinical investigation, numerical values with units, disease scores as a result of an automated
52 calculations, and coding of possible treatment formulas, doses and dosage regimens per
53 disease. **Discussion:** 21 CRFs were designed using OpenClinica v3.14 capturing the 5835
54 coded values per patients. Technical adjustment have been implemented to allow data
55 entry and extraction of this amount of data, rarely achieved in classical eCRFs designs.
56 **Conclusions:** A multidisciplinary endeavour offers complete and harmonized CRFs for AID
57 clinical investigations that are used in TRANSIMMUNOM and will benefit translational
58 research team.

59

60

61 **1 BACKGROUND AND SIGNIFICANCE**

62 Autoimmune and auto-inflammatory diseases (AID) are the third cause of morbidity and
63 mortality in the world ¹. The development of more effective and better tolerated treatments
64 for these chronic and severely disabling diseases is an important public health issue.
65 Recently, genetic studies have highlighted altered biological processes that are common to
66 several AIDs², and others studies have shown that an imbalance between effector T cells
67 and regulatory T cells resulting in the rupture of immune tolerance is associated with AIDs ^{3–}
68 ⁶. This collection of evidence is in line with the proposed reclassification of AIDs to form a
69 continuum of diseases ranging from pure autoimmune to pure inflammatory diseases with a
70 number of diseases displaying variable degrees of both autoimmune and inflammatory
71 disorders ⁷. This is further sustained by immune markers common to several diseases, such
72 as cytokines, which are currently targeted in therapeutics ^{8,9}. The complexity of these
73 diseases, due to the various genetic and environmental factors as well as patient
74 heterogeneity, prompted the scientific community to reconsider research practices with a
75 view to a more integrative approach. In particular, AIDs are associated with multiple and
76 variable immune-related disorders, including dysregulation of the innate immune response
77 or of the adaptive immune response or of both. Systems biology approaches raise the hope
78 that a more comprehensive understanding of cells and tissues in health and disease will
79 open up new avenues for the treatment of patients ^{10,11}. These approaches will transform
80 disease taxonomies from syndromic classification to molecular classification, and their
81 combination, and will allow physicians to select optimal therapeutic regimens for individual
82 patients ¹². Recent studies have successfully identified molecular signatures associated with
83 specific autoimmune diseases ^{4,13–16} as well as in physiological and pathological contexts ^{17–}
84 ¹⁹.

85 Those results led us to setup an observational clinical trial, TRANSIMMUNOM,
86 (NCT02466217) the main goal of which was to revisit the nosology of AIDs through a
87 systems immunology approach. TRANSIMMUNOM participants include patients diagnosed
88 with one out of 19 selected AIDs or one out of 5 control diseases (Figure 1), and healthy
89 donors with no history of autoimmune disorders. The systems immunology approach used a
90 multi-scale deep immunophenotyping on peripheral blood (including transcriptome, TCR
91 repertoire, cytokine expression) and microbiome NGS studies. Importantly, classic routine

92 biology assays as well as clinical investigations are fully part of the data collection strategy.
93 Our aim was to integrate all these data (biology, routine biology and clinical data) so as to
94 allow further cross-analysis of all patients and data to better characterize the immunome of
95 each patient regardless of the initial diagnosis. A similar strategy was initiated by the
96 National Institute of Allergy and Infectious Diseases (NIAID) under the Human Immunology
97 Project Consortium.

98 **2 OBJECTIVES AND OUR CONTRIBUTION**

99 Therefore, we needed to develop data integration approaches to efficiently record and
100 store collected data such that we could easily analyze them afterwards through
101 computational biology approaches. The first challenges of the project were to implement a
102 comprehensive case report form (CRF) covering all diseases in terms of clinical data and
103 biomarkers and to provide a user-friendly, vocabulary-controlled and not expensive
104 platform with standard vocabulary to record all data collected by the clinical assistant during
105 patient interviews. To meet these challenges, we assembled the multidisciplinary “Cohort
106 Management Team (CMT)” composed of clinicians from different specialties, nurses, biology
107 medical doctors, clinical trial methodologists, immunologists and computer scientists.

108 Here we present our electronic CRF (eCRF), designed using an open-source electronic data
109 capture (EDC) tool, capturing more than 5000 multiparametric coded values from 865
110 harmonized clinical and biological parameters per subject included in a multi-disease clinical
111 trial focusing on 24 diseases, 22 areas of clinical investigation and one vast set of routine
112 biology assays. Altogether, we believe that this effort could be of interest for small cohort
113 studies for which the commercially available eCRF services are not accessible.

114 **3 MATERIAL AND METHODS**

115 **3.1 Study population**

116 Patients with one of the following AIDs, of the AID continuum are recruited for
117 TRANSIMMUNOM trial (Figure 1): familial mediterranean fever (FMF), ulcerative colitis,
118 Crohn’s disease, spondyloarthritis, uveitis, myositis (polymyositis, dermatomyositis,
119 inclusion-body myositis, necrotizing and anti-synthetase related myositis), ANCA-related

120 vasculitis (Churg-Strauss' disease and granulomatosis with polyangiitis (ex Wegener), non-
121 ANCA-related vasculitis (such as Behçet's disease, cryoglobulinaemia and Takayasu),
122 rheumatoid arthritis (RA), type-1-diabetes and systemic lupus erythematosus (SLE). We also
123 included patients with diseases exhibiting symptoms similar to those of some AIDs but
124 linked to different gene mutations (control diseases), such as TRAPS and CAPS as a control
125 for FMF, or diseases with a similar autoimmune mechanism with overlapping
126 clinical/biological features, such as antiphospholipid syndrome (APLS) as a control of for SLE,
127 or degenerative diseases that do not have the same mechanism as AIDs such as
128 osteoarthritis for RA and muscular dystrophy for myositis. Finally, healthy volunteers are
129 included.

130 **3.2 Cohort Management Team**

131 Set up to interact with a Clinical Expert Consortium (CEC), a Cohort Management Team
132 (CMT) of biological experts, routine laboratory personnel, clinical trial methodologists and
133 clinical investigation centre harmonized the clinical and laboratory outcomes/results. The
134 CMT ensured that all required data are collected in an appropriate format for analyses and
135 that the questions are unambiguous. The computer scientist defined the data and metadata
136 structure required to minimize non-controlled data entry and to specify the expected
137 values. The overall design was supervised by an immunologist involved in the scientific part
138 of the clinical trial, who liaised between the clinicians and the computer scientist.

139 **3.3 Data collection for eCRF design and coding**

140 Each clinician received an Excel form to be filled in with the description of the item to be
141 recorded in a standardized manner: item ID, item value type (string, decimal); list of pre-
142 determined item values; item value unit (if applicable); item value range (if applicable).
143 Afterwards, all the data collected from the different specialties were grouped and
144 harmonized using CDISC standards.

145 **3.4 OpenClinica implementation**

146 Given the amount of data to be collected across 19 AIDs and 5 control diseases, OpenClinica
147 v3.14, an open-source CFR-part 11 compliant electronic data capture platform has been
148 selected for the design and capture of selected clinico-biological data. A test and production

149 instances have been installed on dedicated and secured CentOS virtual machines with 16Go
150 RAM, 8 cores and 15 Go disk space each. OpenClinica's application server (Tomcat v9.0.6)
151 and database server (PostgreSQL v.9.5) parameters have been upgraded to fit multiple
152 simultaneous data entry and data extraction, in particular JAVA_OPTS for heap memory
153 have been upgraded to 8 Go instead of 1Go Mo by default.

154 **3.5 Patient anonymization**

155 To have completely anonymized subjects who are also unique (no double entries for the
156 same subject in our database because of anonymity) we developed the Anonymized Subject
157 Unification (ASU) system as a completely autonomous system that can be used for any
158 clinical trial. Briefly, ASU takes advantage of a unique identifier of each subject (like Paris
159 Hospital patient number [NIP] or French healthcare registration number [INSEE]) to produce
160 a simple 4-letter code by using a one-way encryption technique.

161 **4 RESULTS**

162 **4.1 OpenClinica as the compromise in designing a multi-disease**

163 **eCRF**

164 The TRANSIMMUNOM observational trial targeted recruitment of 1,000 patients suffering
165 from one out of 24 diseases and healthy controls. During a single visit, patient medical
166 history and clinical investigations are performed together with the collection of samples
167 (blood, serum and feces) for further multi-omics analyses. The goal of the trial is to revisit
168 the nosology of AIDs by defining groups/clusters of patients based on clinical and molecular
169 signatures that cut across disease classification. To deal with the expected amount of
170 heterogeneous (such as disease severity scores, imaging data, biological measures) from
171 routine clinical investigations, and to allow the cross-evaluation of clinical and omics data,
172 we needed to develop an eCRF with a system that allows further omics data integration. We
173 selected OpenClinica (OC) as an electronic data capture (EDC) tool to support our eCRF
174 design. OC is an open-source CRF-part 11 compliant EDC able to design complex eCRFs for
175 large studies^{20,21}. One of the major features of OC is to rely on Clinical Data Acquisition
176 Standards Harmonization (CDASH) from the Clinical Data Interchange Standards Consortium

177 (CDISC)²², which allows the harmonization of clinical and biological data coding. Finally, OC
178 includes the mandatory validation of all recorded data to ensure data quality²⁰. In addition,
179 the main strategy of TRANSIMMUNOM is to cross-analyze data from multiple AIDs, each of
180 which is usually characterized by particular clinical investigation records and biological data
181 measures. We anticipated the final cross-analysis, which would require the same
182 information for all the diseases. Finally, the eCRF had to follow regulatory guidelines and
183 Good Clinical Practices to ensure data entry, traceability and integrity throughout the
184 patient recruitment period. Although installation and implementation of OC is not trivial, as
185 it requires computer science expertise and time, we decided to favour the landscape of
186 possibilities offered by OC to fulfil our study requirement.

187 **4.2 A multidisciplinary workflow ensuring the design of a robust** 188 **multi-disease eCRF**

189 Expertise in different but converging fields was pooled in the CMTs, each of which
190 participated in a 3-step workflow to (1) define the protocol, (2) design and (3) validate the
191 eCRF (Figure 2). The first step of protocol definition involved a Clinical Expert Consortium
192 (CEC) to define the list of items for all the patients with the aim of collecting exactly the
193 same information regardless of the disease. All the clinical specialists together selected a
194 sample of items per specialty so that the CRF was reasonably comprehensive and synthetic.
195 Biology lab experts were also questioned to ensure the feasibility of sample drawing and of
196 the required biology assays. Upon collection and validation of the actual items to be
197 recorded, the specification of the database started with the design of an e-template where
198 the computer scientist structured the information for each item by imposing the format of
199 the data and metadata. Once the e-template was defined, we proceeded to the eCRF
200 design: the CEC, in close collaboration with the computer scientist, designed the clinical
201 coding of clinical investigation data following an unambiguous format for each item with
202 maximized use of a predefined list of responses in order to avoid erroneous data entry.
203 Biology lab experts defined for each parameter measured the value type (string, integer,
204 decimal, Boolean), as well as the unit and range, when applicable. All the information was
205 summarized in a spreadsheet and converted by the computer scientist into a PostgreSQL
206 relational database following the OpenClinica structure. Finally, clinical research technicians

207 evaluated the user-friendliness of the eCRF, the clinical research assistant evaluated the
 208 item relationship constraints, and finally the CMT validated the eCRF with a patient “Zero”
 209 simulation before release for production.

210 4.3 An integrated multi-disease eCRF

211 As AIDs belong to different medical specialties, the CEC comprised clinicians working in
 212 rheumatology, internal medicine, gastroenterology, diabetology, ophthalmology, medical
 213 biology, nephrology and genetics who ensured the feasibility of data collection in terms of
 214 cost, patient morbidity and examination invasiveness. The list of information to be collected
 215 for all the participants was organized in 4 categories. For each recorded item, we defined
 216 the type of value such as free text field, free numerical field, automated calculation, check-
 217 box, drop-down list and calendar/date field (Figure 3). The first group of CRFs was built
 218 under the “Patient description” category and included classic clinical information required
 219 to assess the biology and social environment of the patient. Altogether, we selected 70
 220 items organized as 7 CRFs (Figure 3A & Supplementary material). Each CRF collects 4 to 30
 221 different items. The second set of CRFs focuses on “Common clinical monitoring” and was
 222 organized as 5 CRFs collecting generic clinical data at the day of the visit and accounting in
 223 all for 88 items (Figure 3B & Supplementary material). The third category explore the
 224 “Routine biology monitoring” (Figure 3C, Supplementary material & Table 1) and covered a
 225 wide spectrum of tests.

Hematology	Biochemistry	Protein electrophoresis	Urine analysis	Immunochemistry	Genetic	Serology
Basophils	25-OH Vitamin D	Albuminemia	Creatinuria	ANA	B cell clonality	HIV
Eosinophils	Alkaline phosphatases	Alpha 1 globulin	Hematuria	ANCA	HLAB27	
ESR	ALT	Alpha 2 globulin	Proteinuria	Anti-CCP	HLAB51	
Ferritin	AST	Beta globulin		Anti-dsDNA	HLADR4	
Hematocrite	Calcium	Gamma globulin		Anti-EJ	HLADR8	
Hemoglobin	CH50	Protein electrophoresis peaks		Anti-ENA		
Iron	Cholesterol			Anti-GAD		
Leucocytes	C-Peptide			Anti-HM CR		
Lymphocytes	CPK			Anti-IA2		
MCH	Creatinine			Anti-Jo1		
MCHC	Dyslipidemia			Anti-OJ		
MCV	GGT			Anti-PL12		
Monocytes	Glycemia			Anti-PL7		
Neutrophils	HDL			Anti-SRP		
Platelet	LDL			Anti-ZNT8		
Red blood cells	Phosphate			ASCA		
Transferrin	Triglyrecides			C3		
Transferrin saturation	us-CRP			C4		
				CH50		
				Cryoglobulin		
				Anti-KU		
				Anti-MAD5		
				Anti-MI2		

				Anti-PM/Scl Rheumatoid factor Anti-TIF 1 gamma Anti-RNP		
--	--	--	--	--	--	--

226 **Table 1: List of routine biology assay in the TRANSIMMUNOM trial**

227 Table abbreviation legend: ALT - alanine aminotransferase, ANA - antinuclear antibodies, ANCA - anti-
 228 neutrophil cytoplasmic antibodies, Anti-CCP - anti-cyclic citrullinated peptide antibodies, Anti-dsDNA - Anti-
 229 double stranded DNA antibodies, Anti-EJ - anti-glycyl-transfer RNA synthetase antibodies, Anti-ENA - anti-
 230 extractable nuclear antigens antibodies, Anti-GAD - anti-glutamic acid decarboxylase antibodies, Anti-HMGCR -
 231 anti-3-hydroxy-3-methylglutaryl-coenzyme A reductase antibodies, Anti-IA2 - anti-Islet antigen-2 antibodies,
 232 Anti-Jo1 - anti-histidyl tRNA synthetase antibodies, Anti-Ku - anti- Ku antigen antibodies, Anti-MDA5 -
 233 melanoma differentiation-associated gene 5 antibodies, Anti-MI2 - anti-Mi-2 antibodies, Anti-OJ - anti-
 234 isoleucyl-tRNA synthetase antibodies, Anti-PL7 - anti-threonyl-tRNA synthetase antibodies, Anti-PL12 - anti-
 235 alanyl-tRNA synthetase antibodies, Anti-PM/Scl - anti- nucleolar macromolecular complex PM/Scl, Anti-RNP -
 236 anti-nuclear ribonucleoprotein antibodies, Anti-SRP - anti-signal recognition particle antibodies, Anti-TIF1-
 237 gamma - anti-transcriptional intermediary factor 1-gamma antibodies, Anti-ZnT8 -anti-zinc transporter 8
 238 antibodies, ASCA - anti-*Saccharomyces cerevisiae* antibodies, AST - aspartate aminotransferase, C3 -
 239 complement fraction 3, C4 - complement fraction 4, CH50 - total complement activity, CPK - Creatinine
 240 phosphokinase, ESR - erythrocyte sedimentation rate, GGT - gamma-glutamyl transferase, HDL - high-density
 241 lipoprotein, HIV - Human Immunodeficiency Virus, HLA-B27-B51-DR4-DR8 Human leukocyte antigen -B27-B51-
 242 DR4-DR8, LDL - low-density lipoprotein, MCH - mean corpuscular haemoglobin, MCHC - mean corpuscular
 243 hemoglobin concentration, MCV - mean corpuscular volume, us-CRP - ultrasensitive c-reactive protein.

244

245 These included biological assessment of organ function (liver, kidney, bone marrow) and of
 246 inflammation state and safety, organized in 6 CRF and covering 91 parameters. Finally, the
 247 last set of CRFs recorded “Disease-specific monitoring” data and was subdivided into 3 CRFs
 248 (Figure 3D & Supplementary material) capturing 616 items, including disease activity scores
 249 as described in Table 2. This is thought to be as wide as possible in identifying clinical
 250 parameters not usually collected in a particular disease including imaging and histology
 251 features to allow the identification of disease profile, disease severity and features possibly
 252 shared by diseases. Each clinician of the CEC identified a collection of features observed in
 253 his/her specialty as classic or rare parameters. The CMT gathered all the parameters from
 254 the different specialties and listed them in the clinical status and clinical evaluation CRFs.
 255 Altogether, we selected 865 parameters to describe each patient regardless of the disease.

Disease	Diagnostic criteria	Activity score
Familial Mediterranean Fever/TRAPS/CAPS	Heller ²³ <i>Gene mutation: MEFV; TNFRSF1; NLRP3</i>	AIDAI ²⁴
Ulcerative colitis	clinical and histological features	Mayo ²⁵
Crohn’s disease	clinical and histological features	HBI ²⁶

Spondyloarthritis	ASAS ²⁷ Modified New-York criteria ²⁸	BASDAI ²⁹
Uveitis	non-infectious uveitis	NA
Myositis/dystrophy	clinical and biological	NA
Vasculitis	<u>Behcet's disease</u> ICBD ³⁰ <u>Churg-Strauss</u> ACR ³¹ <u>Cryoglobulinemia</u> ³² <u>Wegener</u> ACR ³³ <u>Takayasu</u> ACR ³⁴	BVAS ³⁵ NIH ³⁶
Rheumatoid Arthritis	ACR; EULAR ³⁷	DAS-28 ³⁹
Osteoarthritis	Kellgren-Lawrence ³⁸	
Type-1-Diabetes	ADA ⁴⁰	IDAA1C ⁴¹
Systemic Erythematosus / Lupus	ACR ^{42,43}	SLEDAI ^{45,46}
Anti-phospholipid syndrome	Sapporo ⁴⁴	

256 **Table 2: Disease specific diagnostic criteria and activity score**

257 **4.4 Clinical coding and CDASH harmonization**

258 Because of the heterogeneity of the selected parameters, clinical coding was designed as an
 259 unambiguous format based on CDASH standards with maximized use of a predefined list of
 260 responses, and was developed as a pragmatic, clinically-validated medical terminology with
 261 an emphasis on ease-of-use data entry, retrieval and data analysis. We therefore defined
 262 and validated for each parameter, wherever possible, the data-type (numerical, text, date,
 263 predefined lists of options, value ranges) and units (when applicable) for all the parameters
 264 identified in order to harmonize the information regardless of the collection time and
 265 person and to avoid errors due to mistyping. Examples are Yes/No check boxes for clinical

266 investigation, numerical values with a list of relevant units according to the parameter,
 267 disease scores as a result of the automated sum of several scores, treatment description
 268 including the coding of possible formulas, doses and dosage regimens (Table 3). We then
 269 coded all the possible/expected values that each item could take and identified 1 to 8
 270 possible variables per item coded as one of the value type. This work was especially critical
 271 for the description of patient treatments. The list of all possible treatments regimen within
 272 each specialty was fully generated with clinicians and is available in the database as a menu
 273 list of 637 variables. Altogether, we built a database with 3815 uniquely coded variables.
 274 However, since clinical status and evaluation of several diseases share identical CRFs, we
 275 reached 5835 possible variables per patient. Altogether, we designed a collection of 21
 276 CDASH harmonized CRFs recording 865 parameters with 5835 coded variables systemically
 277 for all the patients and healthy donors included in the TRANSIMMUNOM trial.

eCRF	Coded question	Coded answer	Value type
Medical History	Medical parameter > AID family history	Presence YES or NO	<input type="checkbox"/>
Hematology	Clinical parameter > Leucocyte count	Numerical (min-max) unit Decimal (4 - 10) 10 ⁹ /L	<input type="text" value="123"/>
Specific activity score	Disease > Ulcerative colitis	Disease score calculation Mayo Score (Sum of 3 scores)	<input checked="" type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
Treatment	Medicine name > Azathiopine	Formula Dosage Posology Tablet 25/50 mg 1/2/3mg/kg/day	<input type="checkbox"/> <input type="text" value="123"/> <input type="text" value="abc"/>

278 **Table 3: Clinical item coding**

279 For 5 exemplary eCRF, a coded question and expected coded answer are described with the type of values to
 280 be entered following Figure 3 legend.

281 5 DISCUSSION

282 Clinical data management is of utmost importance for any clinical study. This includes
 283 clinical information collection, validation and storage, usually completed through the use of
 284 CRFs. While generally clinical research organizations (CROs) propose and use eCRFs, most
 285 academic sponsored clinical studies still take advantage of the cost-benefit of paper versions
 286 of CRFs, sometimes in combination with Excel based databases²⁰. However, such tools,
 287 although convenient, lack validation and data traceability. In addition, they do not usually

288 use harmonized vocabulary and allow free text data entry. These drawbacks were
289 particularly counterproductive in our multi-disease clinical trial from several points of view.
290 First, the main goal of our trial is cross-analysis of multi-omics data obtained with clinical
291 and lab biology data from 1,000 patients selected for one out of 24 AIDs or control diseases.
292 Therefore, we needed an efficient and homogenized set of clinical and routine biology lab
293 for all the patients, which led to the selection of 865 parameters and to the coding of more
294 than 5000 values. This vast amount of data would have been unmanageable using classic
295 paper CRFs and spreadsheets. Second, the amount of information to be collected requires a
296 thorough validation with automated rules and limited free text data entry to avoid
297 mistyping and errors. Again, this cannot be handled using classic methods. Third, for cross-
298 analysis, we need to be able to extract clinical and routine biology data efficiently so that we
299 can filter for parameters as variables of interest (such as gender, BMI, disease activity,
300 autoantibody level). Again, considering the number of patients to be recruited in the
301 TRANSIMMUNOM trial, this would have been impossible. And finally, as regards the disease
302 heterogeneity, it would have been too expensive and complicated to ask an eCRF provider
303 to design such integrated CRFs. For all these reasons, we decided to take advantage of an
304 open-source EDC, Open-Clinica, for the implementation of our eCRFs. Although we
305 anticipated that the design and computer-based requirements would be time-consuming,
306 we found in this tool a number of advantages that allow (i) the integration of a very
307 significant amount of multi-parametric data, (ii) the possibility to design constraints rules
308 between entries to control data entry errors, associated with red flags in the case of errors
309 (for instance a man cannot be pregnant), (iii) the validation of the data entry by a third
310 person who double-checks (the latter advantages being CFR 21 - part 11 compliant) and (iv)
311 the addition of short instructions on the CRF page when needed to guide the data entry and
312 explain to the investigator how to fill in the eCRFs.

313 Altogether, this choice allowed the design of a controlled series of CRFs using harmonized
314 vocabulary to record data across 19 AID patients, 5 control disease patients and healthy
315 donors. This was made possible by the workflow we dedicated to the project, going from
316 the selection of parameters to be collected for all patients regardless of the disease to the
317 coding of all possible values per parameter in a harmonized manner based on CDASH
318 coding. 26 persons were involved in the process, including 14 clinicians, 1 computer
319 scientist, 7 scientists, 3 clinical research technician and assistant as well as 2 medical

320 biologists for more than 100 hours of meetings and discussion over a year and a half. Clinical
321 data coding has the enormous advantages that it (i) pools reported terms in medically
322 meaningful groups, (ii) facilitates identification of common data sets for evaluation of
323 clinical information, (iii) supports consistent retrieval of specific cases or medical conditions
324 from a clinical database and (iv) smooths electronic data interchange of clinical safety
325 information.

326 Finally, our CRFs covers a wide spectrum of clinical and routine biology data of interest for
327 most AIDs, offering the community a pre-designed set of CRFs that can be used together or
328 individually. Although clinical safety was not added to our set of CRFs, because of the non-
329 interventional nature of the TRANSIMMUNOM trial, this could easily be done if needed. This
330 complex set of data has been harmonized and the database designed to store and query
331 efficiently the massive amount of data stored. Altogether, a truly multidisciplinary
332 endeavour led to the design and implementation a collection of 21 CRFs capturing more
333 than 5000 coded values that are now used in TRANSIMMUNOM and could benefit the
334 academic clinical community studying AIDs.

335 **Acknowledgments:** We are grateful to Frédéric Mariotti as an informatic subcontractor for
336 helping in the implementation and maintenance of the OpenClinica instance and to the
337 OpenClinica community, in particular Gerben Rienk Visser from Trial Data Solution for
338 providing help in OpenClinica parameter setup.

339 **Funding:** The work of RL, ID, CR, SH, CA, DK, EMF is funded by the LabEx Transimmunom
340 (ANR-11-IDEX-0004-02) as well as by Assistance Publique-Hôpitaux de Paris and Sorbonne
341 Université. TRiPoD funded the TCR-relevant part of the study.

342 **Author contributions:** RL, ID, CR, CA, AS, DK and EMF composed the Cohort Management
343 Team (CMT) for the design and implementation of the eCRF. RL, ED, PC, AH, BB, DS, FB, GG,
344 PR, OB, KM, JS, PS, MR, CB formed the Clinical Expert Consortium (CEC) and defined the
345 selected clinical and biological data. RL and CR wrote the CRF in agreement with the CMT
346 recommendations. ID performed the computer science part of the work. CM, SH, TO worked
347 on the eCRF validation. EMF coordinated the design and implementation. EMF and RL wrote
348 the manuscript with input from all authors. EMF and DK conceived and supervised the
349 entire work.

350 **Competing interests:** Authors have no competing interests to declare.

351 **6 FIGURES**

		IL-1	IL-6	IL-12	IL-17	IL-23	TNF- α	IFN	Blys	B cell
Autoinflammation	FMF	Red					Red			
	TRAPS/CAPS	Red					Red			
	IBD		Red	Red	Red	Red	Red			
	OA	Red	Red				Red	Red		
	SpA		Red	Red	Red	Red	Red	Red		
	Uveitis	Red	Red		Red	Red	Red	Red		
	Myositis		Red	Red	Red		Red	Red		
	MD						Red	Red		
	Vasculitis							Red		Red
	RA	Red	Red		Red	Red	Red	Red		
	T1D	Red					Red	Red		Red
	SLE	Red	Red	Red	Red		Red	Red	Red	Red
	APLS						Red		Red	Red
			Anti-IL-1	Anti-IL-6	Anti-IL-12	Anti-IL-17	Anti-IL-23	Anti-TNF α	IFN	Anti-Blys

352

353 **Figure 1: TRANSIMMUNOM selected AIDs and control diseases share immune markers and**
 354 **therapeutic strategies**

355 This table shows the list of AIDs selected for the TRANSIMMUNOM trial, belonging to the AID continuum and
 356 their association with cytokines modulation (red) as well as immunotherapies targeting immune markers
 357 (grey). *Abbreviation legend:* Diseases: APLS *anti-phospholipid syndrome*, CAPS *cryopyrin associated periodic*
 358 *syndromes*, FMF *familial mediterranean fever*, IBD - *inflammatory bowel disease*, MD *muscular dystrophy*, OA
 359 *osteoarthritis*, RA *rheumatoid arthritis*, SLE *systemic erythematosus lupus*, SpA *spondyloarthritis*, T1D *type 1*
 360 *diabetes*, TRAPS *tumor necrosis factor receptor-associated periodic syndrome*. Cytokines : IFN *interferon*, IL-1
 361 *interleukin-1*, IL-6 *interleukin-6*, IL-12 *interleukin-12*, IL-17 *interleukin-17*, IL-23 *interleukin-23*, TNF- α *tumor*
 362 *necrosis factor alpha*. Immunotherapies: Anti-BLyS *anti-BLyS monoclonal antibody*, Anti-CD20 *anti-CD20*
 363 *monoclonal antibody* Anti-IL-1 *anti-interleukin-1 monoclonal antibody*, Anti-IL-6 *anti-interleukin-6*
 364 *monoclonal antibody*, Anti-IL-12 *anti-interleukin-12 monoclonal antibody*, Anti-IL-17 *anti-interleukin-17*
 365 *monoclonal antibody*, Anti-IL-23 *anti-interleukin-23 monoclonal antibody*, Anti-TNF α *tumor necrosis factor*
 366 *alpha-blockers*.

367

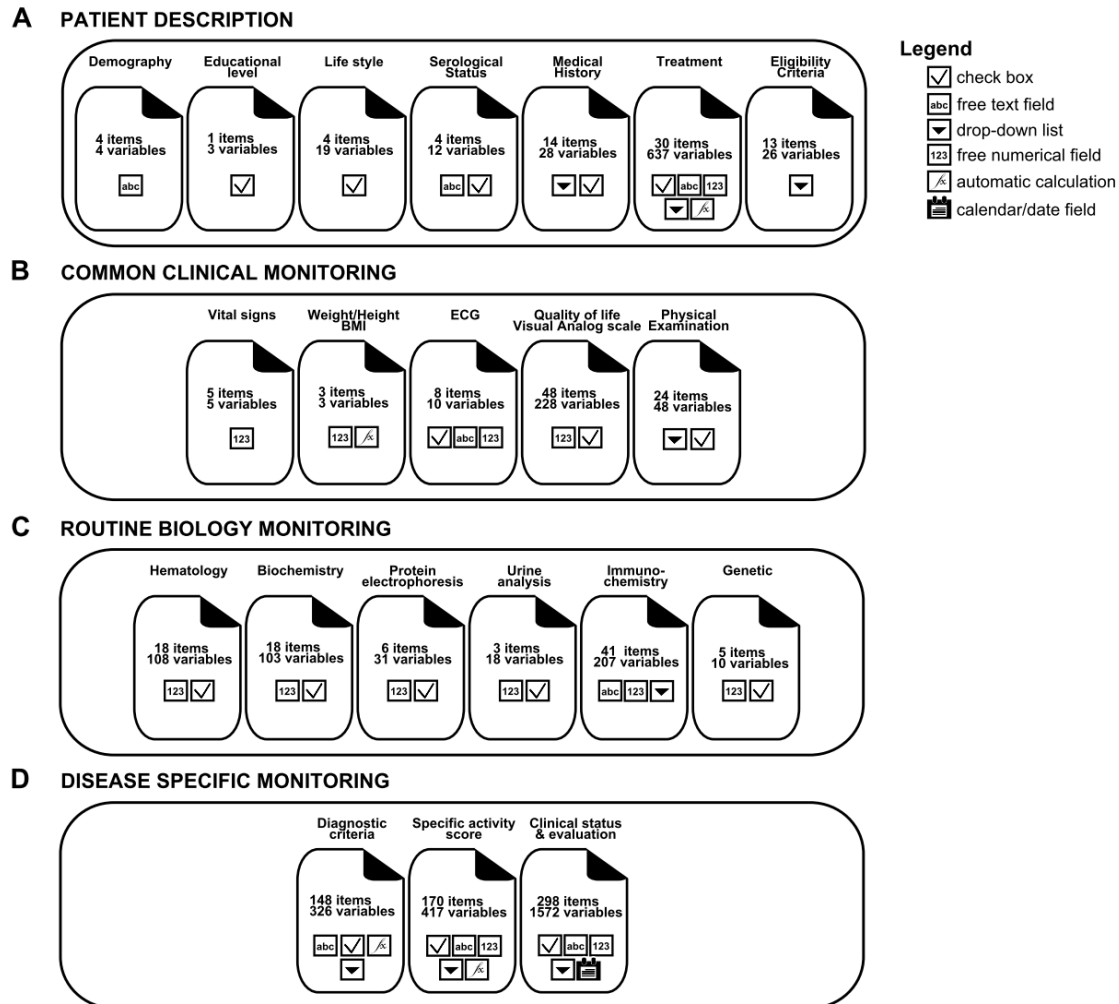


368

369 **Figure 2: eCRF workflow**

370 The figure represents the 3-steps workflow adopted for the eCRF design and implementation: (1) Protocol
 371 design, (2) eCRF design and (3) eCRF validation. In each box are listed the actions, its aim and the person in
 372 charge of it. Color code: Blue: clinical team, purple: biological team, green: computer scientist, and orange: the
 373 trial monitor team and brown the Core management team (see methods for description).

374



375

376

Figure 3: Schematic representation of the TRANSIMMUNOM integrated eCRF

377

Four categories of eCRFs were designed (A-D). Each category is composed of several eCRFs (form icon), each of which contained the indicated number of items for which 1 to 8 variables were coded. The type of values are indicated in the square boxes (see legend), so as to check-box, free text field, drop-down list, free numerical field, automatic calculation and calendar/date field. Altogether, 865 items were coded resulting in 5835 variables organized in 21 eCRFs. See Supplementary material for details on eCRF

378

379

380

381

382

7 REFERENCES

383

1. Getts DR et al. *Immunol. Rev.* **255**, 197–209 (2013).

384

2. Farh KK-H et al. *Nature* **518**, 337–343 (2014).

385

3. Allenbach Y et al. *Am J Pathol* **174**, 989–98. (2009).

386

4. Allenbach Y et al. *PLoS One* **9**, e88788 (2014).

387

5. Buckner JH *Nat Rev Immunol* **10**, 849–59 (2010).

388

6. Rosenzweig M et al. *Curr. Diab. Rep.* **14**, 553 (2014).

389

7. McGonagle D et al. *PLoS Med.* **3**, (2006).

390

8. Magyari L et al. *World J. Orthop.* **5**, 516–536 (2014).

391

9. Moran EM et al. *Clin. Exp. Immunol.* **178**, 405–415 (2014).

392

10. Germain RN et al. *Annu. Rev. Immunol.* **29**, 527–585 (2011).

- 393 11. Pepperkok R et al. *Genome Biol.* **9**, 314 (2008).
394 12. Bielekova B et al. *Front. Neurol.* **5**, (2014).
395 13. Chaussabel D et al. *Nat. Rev. Immunol.* **14**, 271–280 (2014).
396 14. Chaussabel D et al. *Immunity* **29**, 150–164 (2008).
397 15. Saadoun D et al. *N. Engl. J. Med.* **365**, 2067–2077 (2011).
398 16. Terrier B et al. *Arthritis Rheum.* **64**, 2001–2011 (2012).
399 17. Godec J et al. *Immunity* **44**, 194–206 (2016).
400 18. Nehar-Belaid D et al. *J. Immunol.* **196**, 678–690 (2016).
401 19. Querec TD et al. *Nat. Immunol.* **10**, 116–125 (2009).
402 20. Franklin JD et al. *J. Biomed. Inform.* **44**, S103–S108 (2011).
403 21. Leroux H et al. *Stud. Health Technol. Inform.* **168**, 89–95 (2011).
404 22. <http://www.cdisc.org/ICD>
405 23. Livneh A et al. *Arthritis Rheum.* **40**, 1879–1885 (1997).
406 24. Piram M et al. *Ann. Rheum. Dis.* **73**, 2168–2173 (2014).
407 25. Lewis JD et al. *Inflamm. Bowel Dis.* **14**, 1660–1666 (2008).
408 26. Harvey RF et al. *Lancet Lond. Engl.* **1**, 514 (1980).
409 27. Rudwaleit M et al. *Ann. Rheum. Dis.* **68**, 777–783 (2009).
410 28. van der Linden S et al. *Arthritis Rheum.* **27**, 361–368 (1984).
411 29. Garrett S et al. *J. Rheumatol.* **21**, 2286–2291 (1994).
412 30. International Team for the Revision of the International Criteria for Behçet’s Disease
413 (ITR-ICBD) *J. Eur. Acad. Dermatol. Venereol. JEADV* **28**, 338–347 (2014).
414 31. Masi AT et al. *Arthritis Rheum.* **33**, 1094–1100 (1990).
415 32. Quartuccio L et al. *Rheumatol. Oxf. Engl.* **53**, 2209–2213 (2014).
416 33. Leavitt RY et al. *Arthritis Rheum.* **33**, 1101–1107 (1990).
417 34. Arend WP et al. *Arthritis Rheum.* **33**, 1129–1134 (1990).
418 35. Stone JH et al. *Arthritis Rheum.* **44**, 912–920 (2001).
419 36. Kerr GS et al. *Ann. Intern. Med.* **120**, 919–929 (1994).
420 37. Aletaha D et al. *Arthritis Rheum.* **62**, 2569–2581 (2010).
421 38. Kellgren JH et al. *Ann. Rheum. Dis.* **16**, 494–502 (1957).
422 39. van der Heijde DM et al. *Ann. Rheum. Dis.* **49**, 916–920 (1990).
423 40. American Diabetes Association *Diabetes Care* **35 Suppl 1**, S11-63 (2012).
424 41. Mortensen HB et al. *Diabetes Care* **32**, 1384–1390 (2009).
425 42. Petri M *Rheum. Dis. Clin. North Am.* **31**, 245–254, vi (2005).
426 43. Hochberg MC *Arthritis Rheum.* **40**, 1725 (1997).
427 44. Wilson WA et al. *Lupus* **10**, 457–460 (2001).
428 45. Bombardier C et al. *Arthritis Rheum.* **35**, 630–640 (1992).
429 46. Petri M et al. *Lupus* **8**, 685–691 (1999).
430