Reference: PhUSE Data Transparency Working Group's Comments on Health Canada's Draft Guidance for the Implementation of the Public Release of Clinical Information of 10 April 2018.

Attention: hc.rmod.stakeholders-intervenants.dgro.sc@canada.ca

Copenhagen, 25 June 2018

We would like to thank Health Canada for the opportunity to review this important guidance. Please find our comments and questions in the table below. We hope they are useful.

PhUSE is an independent, not-for-profit organisation run by volunteers. Since its inception, PhUSE has expanded from its roots as a conference for European statistical programmers to a global membership organisation and platform for the discussion of topics encompassing the work of data managers, biostatisticians, statistical programmers and eClinical IT professionals.

PhUSE launched a Data Transparency Working Group in 2014, gathering a cross-industry team including pharmaceuticals, CROs, software professionals, data anonymization experts and academics to define data anonymization standards and address key aspects of data sharing. The Working Group released, in particular, the PhUSE Data De-Identification Standard for SDTM 3.2, which is used by numerous organisations to share data and referenced in important regulatory and industry guidance such as the EMA External Guidance on the Publication of Clinical Data, the Ontario De-Identification Guidelines for Structure Data and the TransCelerate Data De-Identification Model Approach.

PhUSE is also a stakeholder of the FDA, EMA and PMDA and hopes there will be other opportunities to assist Health Canada with relevant regulations or guidance where its expertise can add value.

On behalf of the PhUSE Data Transparency Working Group,

Yours truly,

Jean-Marc Ferran
PhUSE Data Transparency Working Group Lead
Email: jean-marc.ferran@phuse.eu

Line	Section	Author	Comment
148	1.0	Sarah Nevitt (University of Liverpool)	Within the definitions of 'Directly identifying Variable' (line 148) and 'Indirectly identifying Variable' (line 151), the analytic usefulness of these types of variables are dichotomised as 'not useful' and 'useful' respectively without any further description. The usefulness of a variable is discussed more on page 18 of the guidance, which provides context for these subjective judgements. For the purpose of accurate definitions, we suggest that the reference to whether the variable is analytically useful or not be removed from the definitions. In addition to the lack of context within these definitions, the 'usefulness' of a variable does not define whether it is directly, indirectly or not identifying.
245	2.3	Sarah Nevitt (University of Liverpool)	It is stated that "The decision to disclose clinical information within interim analyses will be taken case by case", according to the table of considerations, and, later in a subsequent section (line 408), that "Health Canada retains final decision on what information is publicly released". Will the decision regarding release of interim analyses be made solely by Health Canada, or will the manufacturers, particularly those involved in the design, analysis and interpretation of the interim results – who will best understand the likely impact – be involved in this decision?
General		Sarah Nevitt (University of Liverpool)	General question regarding past submissions: will a date limit be placed on how far back requests can be made? Or will the feasibility of requests for information from past submissions be judged on other means, such as resources required to prepare information (e.g. line 308-9 refers to digitization of paper records).
574-587	Appendix A -C	Sarah Nevitt (University of Liverpool)	I would also like to add that, despite some issues (see next comments) in understanding some of the content of Appendix A to C, I like the concept of listing what will not be publicly released in terms of sections of the documents in question. From my 'researcher'

Line	Section	Author	Comment
			(data user) perspective, this is very helpful for knowing exactly what is available from this source, and where I may need to request additional information (such as IPD).
574	5 (Appendix A, B and C general comments)	Sarah Nevitt (University of Liverpool)	The relevance of the CBI columns is not clear in these two appendices as everything is classified as 'Not CBI'. Furthermore, this column potentially contradicts Section 4, where Health Canada may consider information such as secondary or exploratory outcome measure data or end points (line 443-450) as CBI. The final column refers to whether specific sections of clinical summaries, clinical overviews and clinical study reports will be released. The title of this column is 'Public Proactive Release'. Will the same approach be taken for requests for past submissions?

Line	Section	Author	Comment
578	Appendix B	Sarah Nevitt (University of Liverpool)	Presumably, the proactive release of certain sections listed here will only be done following appropriate anonymization/redaction? We note that 16.2 Patient Listings will not be proactively released but within the following sections which are marked yes for proactive release. It is likely that individual patient data for some or all patients will be present: ○ 11.4.3 – Tabulation of Individual Response Data ○ 11.4.6 – By-Patient Displays ○ 12.2.4 – Listing of Adverse Events by Patient ○ 12.3.2 – Narratives of Deaths, Other Serious Adverse Events and Certain Other Significant Adverse Events ○ 12.4.2.2 – Individual Patient Changes ○ 12.4.2.3 – Individual Clinically Significant Abnormalities ○ 14.3.3 – Narratives of Deaths, Other Serious and Certain Other Significant Adverse Events ○ 14.3.4 – Abnormal Laboratory Value Listing (Each Patient). We suggest highlighting in the guidance that these sections that are to include patient data and special considerations with regards to anonymization must be undertaken as a result.

Line	Section	Author	Comment
155	1.2	Luk Arbuckle (Privacy Analytics)	The terms anonymization and de-identification are often used interchangeably, yet the definition defines anonymization as a process that renders data de-identified (and no definition is provided for de-identified). Also, the definition of anonymization that is proposed includes process and conditions, some of which could be misunderstood without proper context. • "stripped of direct identifiers" could be misinterpreted to mean deletion, which would not allow for linking via the use of (irreversible) pseudonyms. • "code" could be misinterpreted to mean any linkage built from the structure of data that the trial sponsor or Health Canada will have (as custodians of the identifiable data). • A risk threshold is dependent on context and is usually described outside of a definition so that it may be more generally used. Recommendation: Use a more generic definition of anonymization. • ISO/TS 25237: A process that removes the association between the identifying data and the data subject. • IPC Ontario: The process of removing personal information from a record or data set. • Other: methods for reducing re-identification risk, usually based on restricting the amount of, or transforming, the data shared or released, and applying mitigating controls in the environment in which the data are shared.
166	1.2	Luk Arbuckle (Privacy Analytics)	Re-synthesis can also be applied to suppressed values, by imputing another value (e.g. entirely at random, an average value). It seems that the term refers to the HIPS (Hide In Plain Sight) method.

Line	Section	Author	Comment
			Recommendation: Generalize the definition to include suppressed values that are re-synthesized, e.g. "Means a technique to randomly generate values that are substituted for the original values". And refer to HIPS, which is more commonly known by the community.
169	1.2	Luk Arbuckle (Privacy Analytics)	Randomization generally means to reorder following a distribution (in this context it would mean every value has equal probability of being selected in the process of reordering). In the case of direct identifiers, the original values would not be used; in the case of indirect identifiers, the original values may be included using a technique from statistical disclosure control known as "data swapping". What is being described sounds like perturbation. Recommendation: replace "randomization" with "perturbation", and include a new definition of randomization (and specify if this is in the context of direct identifiers, in which case the original values should not be used, instead fake ones or from gazetteers).
503	5.2	Luk Arbuckle (Privacy Analytics)	Date of birth is classified as indirectly identifying in all other guidance, including PhUSE. It is replicable, distinguishable, and knowable. It is also analytically useful. However, depending on the population and accompanying set of indirectly identifying variables, it may be generalized (e.g. year of birth) or suppressed for some individuals. Recommendation: replace with another example, such as name (or initials). Many names (and certainly initials) are not uniquely identifying (e.g. John Smith), but they are not analytically useful.
504	5.2	Luk Arbuckle (Privacy Analytics)	Analytic utility, as described in other guidance, refers to the ability to perform mathematical or statistical computations on the values in a variable (e.g. mean, regression). Subject identification numbers may be useful to preserve "referential integrity" (i.e. the ability to link records or information to a single subject), but are not used to calculate statistics.

Line	Section	Author	Comment
			Recommendation: explain what analytically useful means, and remove "with limited exceptions" (perhaps explain referential integrity, how it is an important property to maintain through pseudonmization but not related to analytic utility).
515	5.2	Luk Arbuckle (Privacy Analytics)	The sentence could be interpreted to mean variables alone, and not in combination with one another or external sources of information (e.g. date of birth alone instead of date of birth combined with postal code, or how that can be combined to an identity from an external database). Recommendation: include "alone or in combination with other information" to the sentence, e.g. "Variables which do not present a serious risk of re- identifying an individual, alone or in combination with other information…".
524	5.2	Luk Arbuckle (Privacy Analytics)	The standards on direct identifiers are different in unstructured text data than they are in structured data. In unstructured data there is the possibility of leaking a direct identifier, hopefully hidden in plain sight through re-synthesis, and this leak rate should be kept to a minimum. Recommendation: suggest the measurement of re-identification risk associated with potential leaks of direct identifiers.
545	5.2	Luk Arbuckle (Privacy Analytics)	The re-identification risk suggested is a standard for public health reporting; however, for more sensitive information stricter thresholds have been applied. Recommendation: where the sensitivity of the data being released is greater, a lower re-identification risk threshold could be considered. This is discussed in the literature in, for example, "A unified framework for evaluating the risk of re-identification of text de-identification tools": https://www.sciencedirect.com/science/article/pii/S1532046416300697.

Line	Section	Author	Comment
559	5.2	Luk Arbuckle (Privacy Analytics)	As mentioned in the comment for line 504, analytic utility typically refers to the ability to perform mathematical or statistical computations on the values in a variable, whereas referential integrity refers to the ability to link records or information across records. Recommendation: replace analytical utility with referential integrity, e.g. "other directly identifying variables which are needed to maintain referential integrity…".
563	5.2	Luk Arbuckle (Privacy Analytics)	The EMA currently does not require a risk-based approach or a re-identification risk measurement for the sponsor's/manufacturer's staff, in which there is more likely to be a zip/postal code for place of work), but instead allows for a rules-based approach. In the table below 562, it refers to signatures, job titles/positions, etc. that are usually related to sponsors' staff rather than patients. Recommendation: if a risk-based approach to staff information (e.g. zip/postal code at place of work) is intended, this should be made clear throughout the guidance.
566	5.2	Luk Arbuckle (Privacy Analytics)	Geographical region is considered indirectly identifying in all other guidance, and can be generalized on many levels depending on re-identification risk. PhUSE guidance for EMA Policy 0070 recommends that country be generalized to continent, in order to manage risk and because country is not considered to be of high analytic utility. It is not uncommon in international trials to have only a very small number of trial participants in a single country, increasing re-identification risk dramatically. Recommendation: explain why "country should remain unmodified", and consider removing this from the guidance (thereby allowing country to be generalized in the same way as all other indirectly identifying variables).
3.1	348-350	Alexandra Marquart (Boehringer Ingelheim)	"Upon receipt of a request for clinical information, Health Canada will prioritize the request (see below), conduct an internal search for records, and publish the requested information on its clinical information portal."

Line	Section	Author	Comment
			Does Health Canada plan to consult with the sponsor before publicly posting past submissions? Will the sponsor be notified about a request at Health Canada so the sponsor can provide input? Please clarify. Will there be a consultation with the sponsor before publicly posting past submissions? Or will there be the possibility to be notified so the sponsor can provide input?
2.0	214-216	Alexandra Marquart (Boehringer Ingelheim)	"the process through which Health Canada will protect personal information and information that continues to be CBI": Please clarify whether Health Canada will be responsible for protecting personal information and CBI for past submissions. Will Health Canada use redaction to protect that information or another method? Who will assume responsibility in case of past submissions for redaction? Will it be Health Canada?
3.1	348-350	Alexandra Marquart (Boehringer Ingelheim)	"Upon receipt of a request for clinical information, Health Canada will prioritize the request (see below), conduct an internal search for records, and publish the requested information on its clinical information portal." Does Health Canada plan to consult with the sponsor before publicly posting past submissions? Will the sponsor be notified about a request at Health Canada so the sponsor can provide input? Please clarify. Will there be a consultation with the sponsor before publicly posting past submissions? Or will there be the possibility to be notified so the sponsor can provide input?
3.3	408	Kim Musgrave (Amgen)	"Health Canada retains final decision on what information is publicly released." However, 3.4 states the manufacturer will provide the final versions so they will remain as the data controller. Following Health Canada's review, the manufacturer must submit a final version of the redacted documents, according to Health Canada instruction.

Line	Section	Author	Comment
			Could you clarify who is the data controller? Will Health Canada then share this responsibility with the sponsor?
5.2	535	Jean-Marc Ferran (Qualiance)	The section on "Reference Population" is a great addition compared to other data anonymization guidance as it clearly itemizes different options for companies to consider when modelling the risk.
General		Jean-Marc Ferran (Qualiance)	Are there fundamental differences between the terms "analytic utility" and "CBI" used in Health Canada guidance and "data utility" and "CCI" used in EMA Policy 0070 guidance, respectively? If that is not the case, would an alignment in terminology be considered? Otherwise, could you please explain how you consider these terms to differ?