

In the Bookstores

Guide to the De-Identification of Personal Health Information

Written by Khaled El Emam
ISBN-10: 466579064
Publisher: Auerbach Publications (T&F)
414 pages
Paperback: £42.66

Reviewed by Alison McIntosh
AAG Medical Writing

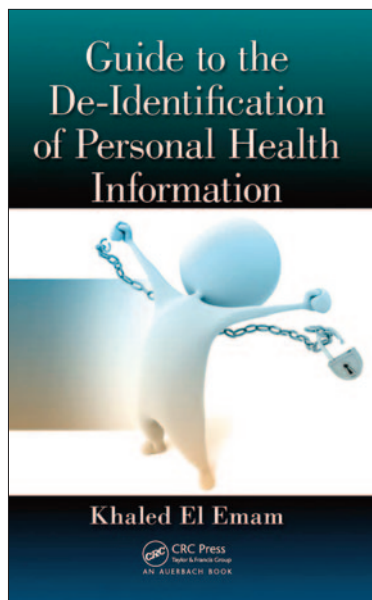
doi: 10.56012/hbvt6966

Increasingly, regulatory authorities require sponsors to publish clinical documents to increase transparency and openness and improve public trust in clinical research. In doing so, the sponsor has a responsibility to ensure that personal data that might allow individuals to be identified is not released in the published version.

Alongside regulatory public disclosure, de-identified clinical data can be made available for secondary health-related research purposes. A third party may submit a request to a sponsor to perform secondary analysis or research of patient clinical trial datasets. Again, prior to release for secondary purposes the sponsor must ensure that the identity of a trial participant is not inadvertently revealed. The amount of de-identification required to ensure anonymisation then has to be balanced against preserving data utility for the researcher.

Regulatory medical writers have an important role to play in the process of preparing clinical documents suitable for public disclosure, and separately can also have involvement in the process of preparing clinical trial datasets suitable for data sharing. There is overlap in the methodology used to manage the risk of de-identification of individuals in both processes.

Guide to the De-Identification of Personal Health Information provides information on the different methodologies that can be employed to manage the risk of reidentification of personal health information. The author, Professor El Emam, is a true expert in the field and is a member of the European Medicines Agency Technical Anonymization Group (<https://www.ema.europa.eu/en/human-regulatory-overview/marketing-authorisation/clinical-data-publication/technical-anonymisation-group>). Over many years, he has been influential in



advocating and developing practical methodology and software tools that allow health data to be accessible whilst importantly maintaining patient privacy. He also promotes the idea that when properly applied, de-identification of personal health information can allow the use of health data for important secondary reasons, including health-related research.

For those working in the transparency and disclosure arena this book is a practical guide designed to provide “a valuable and much needed resource for all data custodians who use or disclose personal health information for secondary purposes.” Professor El Emam defines secondary purposes as “non-direct care use of personal health information including, but not limited to, analysis, research, quality/safety measurement, public health, payment, provider certification or accreditation, and marketing and other business including strictly commercial activities.”

The book has 28 chapters and hence can be a little daunting at first view. However, to help the reader it is organised into four main sections covering: 1. The case for de-identifying personal health information (Chapters 2-9); 2. Under-

SECTION EDITORS



Alison McIntosh
aagmedicalwriting@gmail.com



Pablo Izquierdo
pablo.izquierdo@ppd.com

standing disclosure risks (Chapters 10-15); 3. Measuring re-identification risk (Chapters 16-19); and 4. Practical methods for de-identification (Chapters 20-21). A fifth section entitled “End Matter” provides appendices with supporting materials (Chapters 22-28).

Although the book is written principally with disclosure of personal health information for secondary purposes as the main thrust, there is much that is useful to those medical writers working in the regulatory transparency and disclosure arena. Early chapters provide a useful background to the topic and the later chapters provide in-depth methods for managing reidentification risk.

A general background to the Health Insurance Portability and Accountability Act (HIPAA) is provided together with a discussion around HIPAA Safe Harbor principles. The author argues that the Safe Harbour approach, upon which much of the de-identification methodologies employed in regulatory public disclosure are currently based, “does not provide adequate protection”. Rather, he recommends the statistical method as a better risk-based de-identification method to apply (Chapter 1). The chapter entitled “Scope, Terminology and Definitions” provides an introduction to

identifiers and their classification. Applicable rules to help determine if a variable is a direct identifier or a quasi-identifier are presented and illustrated in a useful in-text table (Chapter 10).

Although as medical writers we do not need a full understanding of the statistical concepts behind data anonymisation, it is useful for us to have a working knowledge of them. A thorough review of the types of statistical approaches that can be implemented together with measuring the probability of re-

identification risks are covered in later chapters.

While quite complicated, explanations of how data utility loss can be mitigated and risk assessed, together with any limitations of the approaches being used, will be of help to medical writers working as a subject matter expert (SME) in the transparency and disclosure arena. Taken

This book is a valuable and much needed resource for all data custodians who use or disclose personal health information for secondary purposes.

together they offer extra insight on how best to achieve de-identification of personal health information using risk-based approaches.

The general summary of commonly used and recommended data masking includes descriptions of suppression, randomisation, irreversible and reversible coding (Chapter 14). Explanations of why other techniques are generally not to be recommended are also presented, including constraining names, adding noise, character scrambling, character masking, truncation, and encoding. For non-statisticians this chapter provides detailed explanations of statistical terminology that, in my experience, is not always well-defined. Specifically, this chapter will allow

medical writers working as SMEs in this area to have a better grasp of the approaches that statistical colleagues implement when a quantitative anonymisation strategy is used.

El Emam suggests that generalisation (e.g., generalising date of birth to a five-year age interval) and suppression (e.g., removing a patient or a visit from the data set) methods “have the most acceptability” and both are discussed in much detail in Chapter 20 “De-identification Methods”.

Although this book was first published in 2013, even now there is no single method recommended for de-identification of health information. The book provides an in-depth and detailed background to the statistical concepts

that can be applied to de-identify clinical datasets. As you grapple with preparing clinical documents suitable for public disclosure, my personal view is that although this book is quite technical, and may not answer all the questions you have about data anonymisation, it provides you with an appreciation of the complex methodologies involved.

Professor El Emam has published more recent books and journal articles. A full list of his publications, including book titles, can be found on the Electronic Health Information Laboratory website: <https://www.ehealthinformation.ca/publications> where you can also sign up to receive monthly newsletter updates.

A vibrant, multi-colored banner for the EMWA's 58th Conference. The background is a collage of overlapping geometric shapes in shades of green, purple, orange, red, and blue. The text is arranged in a central, layered fashion. At the top, it reads "VIRTUAL CONFERENCE WITH REGIONAL HUBS • VIRTUAL CONFERENCE WITH REGIONAL HUBS". Below this, the main title "EMWA's 58th Conference" is displayed in large, bold, white letters, with "NOVEMBER 2024" underneath in yellow. A white horizontal line separates the title from the text "Check emwa.org for updates about workshops, speakers, and regional hubs!". Another white horizontal line follows. Below that, it says "VIRTUAL CONFERENCE WITH REGIONAL HUBS" in yellow, and "Registration is now open!" in white. At the bottom, the text "VIRTUAL CONFERENCE WITH REGIONAL HUBS • VIRTUAL CONFERENCE WITH REGIONAL HUBS" is repeated in white, mirroring the top. The words "VIRTUAL CONFERENCE WITH REGIONAL HUBS" are also written vertically on the left and right sides of the banner.