ECCW Shung 宋朱彩雲

Data cleaning, maintenance, and analysis of the HA SARS Collaborative Groups database

Key Messages

1. Data from various operational systems were integrated into a structured data collection system enabling the construction of a comprehensive, timely, complete, and accurate HA SARS Collaborative Groups dataset.
2. This dataset facilitated further in-depth, sophisticated analyses on the risk/prognostic factors by comparing outcomes of different treatments at different time points of disease progression.
3. Inputs from clinicians on data validation and management, data analyses, and modelling helped to better understand the disease.
4. The experience gained has laid foundations for conducting clinical and health services research studies, should future epidemics of infectious disease affect Hong Kong.

Introduction

During the 2003 severe acute respiratory syndrome (SARS) outbreak, 8000 patients worldwide were infected, resulting in nearly 800 deaths. Of these, 1755 patients were infected in Hong Kong, of whom nearly 300 died. The causative agent was soon identified as a novel coronavirus. Scientific knowledge on and clinical experience with SARS grew rapidly. Initially, diagnosis was based on a constellation of clinical and epidemiological findings, and later such features were backed up by laboratory confirmation. Many studies/reviews on various aspects of SARS are of limited value by their sample size, inclusion of patients based purely on clinical grounds, and by the use of early outcome indicators (such as the day-21 mortality only) that were not representative of the eventual outcome group.

Exploratory work on the clinical data, risk and prognostic factors, correlations and modelling were conducted by staff in the Hospital Authority (HA) Head Office with professional and expert advice from the clinicians. Published studies from the respective hospitals/academics were referred to.

Aims and objectives

The main objectives were: (1) to collate and manage data from various sources and consolidate it into a comprehensive HA SARS Collaborative Groups (HASCOG) database for 1755 SARS patients encountered in Hong Kong; and (2) to support research studies and analyses on SARS patients.

Methods

As HA public hospitals treated all local SARS cases, each patient was identifiable via a unique identification number, which could be used to link many clinical parameters via existing HA systems. They included details of clinical management and the eventual outcome. According to the usual HA practice, daily clinical and other vital data for SARS cases were recorded manually in medical records. These data were abstracted by frontline staff in the respective hospitals and entered into the main database by HA Head Office staff. In addition to relevant research nurses, information technology and statistical staff, and one clinician dedicated to the task reviewed the data for completeness and validated it on an ongoing basis.

The HASCOG dataset was amalgamated from different sources. The HA Clinical Management System (CMS) provided demographics, admission and discharge, pharmacy records, and laboratory results. Symptom-onset dates, contact history, and presenting symptoms were collected through a real-time case questionnaire survey conducted by the Hong Kong SAR Government Department of Health. External data on polymerase chain reaction (PCR) and serology findings were incorporated from various laboratories in the Department of Health, the University of Hong Kong, the Chinese University of Hong Kong, and the HA. Clinical data on co-morbidities, daily vital signs and details of drug treatment, oxygen and ventilatory therapies were manually abstracted from the medical records by clinical staff using a standardised data entry form. Defined

Hong Kong Med J 2008;14(Suppl 1):S11-3

Statistics and Research Unit, Hospital Authority Head Office, 147B Argyle Street, Kowloon, Hong Kong SAR, China
ECCW Shung

RFCID project number: HA-CS-001

Principal applicant and corresponding author:
Mrs ECCW Shung
Statistics and Research Section, Hospital Authority Head Office, 147B Argyle Street, Kowloon, Hong Kong SAR, China
Tel: (852) 2300 6371
Fax: (852) 2895 2167
E-mail: shungccw@ha.org.hk

co-morbidities included: chronic obstructive pulmonary disease, cardiovascular disease, cerebrovascular accident, active cancer, diabetes mellitus, chronic renal insufficiency, and chronic liver disease. Chest radiographs were scored retrospectively by blinded radiologists according to a five-point scale for each of the six lung zones. Radiographic scorings were confined to five milestones: (1) at presentation, (2) upon commencement of ribavirin treatment, if any, (3) upon commencement of pulse steroid treatment, if any, (4) at peak lung opacification, and (5) the last film prior to death/discharge. Laboratory findings were recorded according to a common reference point or scale or else as absolute values.

After establishing the database, it was 'cleaned'. This included tracking and retrieval of missing episodes and clinical data, double checking of dubious results/details with the frontline staff or checking them against the CMS records. Where necessary, the clinical data were cross-checked with in-charge clinicians and/or with patients' manual medical records. Excel files on serology and PCR findings from external laboratories, which had been prepared using different data formats, were painstakingly standardised and entered into the records. Serology tests were also requested post-mortem to obtain laboratory confirmation of suspected SARS cases. After investigation and interpretation of every essential piece of data from different sources, further rectification and confirmation with the data source was carried out, if necessary. Thus, results of tests arranged externally by collaborating academic institutions and the Department of Health were tracked. Repeat serology tests were conducted if initial results appeared inconsistent or dubious. Causes of death were confirmed by reviewing the medical records, as were the diagnoses and data entered in the discharge summaries. All patients were followed up retrospectively until death or hospital discharge. Two specific adverse outcomes were studied: hospital mortality and oxygenation failure. The latter was defined by a PaO_2/FiO_2 (P/F) ratio of less than 200 mm Hg, the level used to define acute respiratory distress syndrome.

Using different methods of statistical analysis and modelling, this collection of observational data was retrospectively studied. Observations, interpretation, and modelling results were discussed with clinicians, who provided advice or feedback based on their clinical knowledge, expertise, and intuition.

Results

A comprehensive, cleaned, and managed database on the demographic, epidemiological, clinical, laboratory, and radiological profile of the 1755 local SARS patients was constructed to facilitate various studies and reviews. Several manuscripts were then prepared that incorporated data from this HASCOG database.[1-5]

Apart from the published reports using information from the cleaned HASCOG database, the results have been presented in numerous HASCOG meetings, as well as local and international fora, conferences, and other internal meetings. These events provided an opportunity for health professionals to exchange valuable experience and share significant research findings on various aspects of SARS.

Discussion

Our project confirms that retrospective collection of clinical data on a community-wide epidemic, such as the 2003 SARS epidemic, can yield good quality data, if there is careful pre-planning and coordination. Examples of clinical parameters that required cleaning are considered below.

Symptom onset date

Symptom onset date (SOD) plays an important role in facilitating the understanding of the clinical course and transmission dynamics of an infectious disease. 'Days from SOD' has been used as the common reference point for profiling and analysing clinical and investigative data, in order to eliminate inter-patient differences in the timing of their presentation. Regrettably, a substantial number of patients with a designated SOD on the HA database differed from that recorded by the Department of Health. Elaborate attempts were then made to review patient clinical records to ascertain, as much as possible, the true SOD of such cases.

Presenting symptom complex

In the initial HASCOG data collection, the presenting symptom complex was based on each patient's admission notes, written by the admitting doctor. During data verification, it was noted that recording of symptoms by the admitting doctor had not been entered consistently. At some hospitals, data were entered into a standard table listing all possible symptoms. Other hospitals relied on the admitting doctor's brief clinical notes, in which case failure to mention a symptom was interpreted as absence of that particular symptom.

Serum enzymes

Prior to our data cleaning exercise, a few studies suggested the importance of certain serum enzyme levels (such as for lactate dehydrogenase and creatine kinase) as being prognostically important. After the data cleaning exercise, it was realised that there were major differences in normal ranges among various hospital laboratories, which greatly influenced the interpretation of the laboratory values. In particular, for the same laboratory test on enzymes, different reference ranges were adopted by different laboratories. Therefore a ratio of the measured value over the upper limit of the respective reference range was used to enable more appropriate inter-laboratory comparisons.

PaO_2/FiO_2 ratio

This ratio is an internationally recognised surrogate measurement of the efficiency of oxygenation by the

lungs, and has been used in the international literature to define stages of acute lung injury. To ascertain this ratio, one needs to know the simultaneous PaO_2 and the FiO_2. Although in the initial HASCOG database, FiO_2 had been entered, the exact mode of oxygen delivery was not always recorded. Based on such details, it was possible to derive FiO_2 equivalence values. Furthermore, information on the various types of oxygen masks used was not available. This made it very difficult to reliably estimate the FiO_2 for very sick patients, as they could have been using a simple mask (with up to 50% FiO_2), a Hudson mask (with up to 80% FiO_2), or a non–re-breathing mask (with 100% FiO_2). These uncertainties all tended to compromise the quality of the HASCOG data on gas exchange and the severity of respiratory embarrassment.

Co-morbidities

The HASCOG data might have been all-embracing in terms of data on patient co-morbidities, but there were no standard definitions for many of them. For instance, how does one define chronic obstructive pulmonary disease in a patient? A set of clinical definitions for each of the co-morbidities was therefore prepared, and searched for in the clinical notes for data verification in the respective category.

Based on this experience handling and managing the HASCOG database, it is evident that a well pre-meditated organisation of data to be collected could greatly enhance the accuracy and reliability of the information collected. Pre-meditated organisation of the data collection entails uncompromised attention to completeness of the clinical dataset, standardised and precise definitions on essential clinical information, and standardised data collection formats.

The experience gained by HA staff in organising the HASCOG database can enhance understanding of the type of information that could be important in the next epidemic of an infectious respiratory disease. The necessary preparatory work for a possible future epidemic of avian influenza (H5N1) is now in place, through the establishment of the Avian Influenza Collaborative Group (AICOG) in 2006. A multi-centre, double-blinded, randomised controlled trial on the efficacy and safety of high dose versus WHO-recommended dose of oseltamivir in the treatment of avian influenza (H5N1) has been supported. The need to record specific clinical data at respective time points after symptom onset and a particular format for recording have therefore been proposed and agreed based on this work.

Conclusions

The SARS dataset referred to here is the most comprehensive and accurate data on the 1755 SARS patients in Hong Kong available to date. Interpretation, standardisation, and validation of the data collected from different sources (including external sources), either manually or downloaded from HA clinical system, is a tedious and laborious task. A dedicated team or expert group, with suitable clinical input, should be formed to explore the data, to review discrepancies, and to propose and agree on rules for its validation, management, and dissemination. The experience gained by the HA in organising the HASCOG database has provided the necessary expertise in the further understanding of SARS and possible post-SARS events. These findings enable better understanding of clinical profiles, disease progression, and associations with risk factors and interventions with the patient outcomes. These types of studies can help enrich clinical knowledge and understanding in these areas.

Recommendations

Using the HASCOG database as an example, advance planning and consensus on the data requirements and formats for its collection could enhance the quality and efficiency of future endeavours. The experience gained in this exercise was invaluable and sheds light on strategies for data collection in the event that a similar infectious respiratory disease emerges in the future.

Acknowledgements

This project forms part of a series of studies commissioned by the Food and Health Bureau of the Hong Kong SAR Government and funded by the Research Fund for the Control of Infectious Diseases (Project No. HA-CS-001).

References

1. Chan JC, Tsui EL, Wong VC; Hospital Authority SARS Collaborative Group. Prognostication in severe acute respiratory syndrome: a retrospective time-course analysis of 1312 laboratory-confirmed patients in Hong Kong. Respirology 2007;12:531-42.
2. Malik Peiris JS, Tsang DNC, Lim WWL. Virological diagnosis of SARS. In: Chan JCK, Taam Wong VCW, editors. Challenges of severe acute respiratory syndrome. Singapore: Elsevier; 2006:299-316.
3. Yam LY, Lau AC, Lai FY, et al. Corticosteroid treatment of severe acute respiratory syndrome in Hong Kong. J Infect 2007;54:28-39.
4. Yam LY, Chan AY, Cheung TM, et al. Non-invasive versus invasive mechanical ventilation for respiratory failure in severe acute respiratory syndrome. Chin Med J (Engl) 2005;118:1413-21.
5. Antonio GE, Ooi CG, Wong KT, et al. Radiographic-clinical correlation in severe acute respiratory syndrome: study of 1373 patients in Hong Kong. Radiology 2005;237:1081-90.