The Labour and Welfare Bureau of the Government of the Hong Kong Special Administrative Region

Provision of Consultancy Services for Developing a Central Databank on Children
(Ref: LWB CoC/7-2/1/2)

Final Report

August 2023

Table of Contents

1. Introduction 1
1.1. Project Overview 1
1.2. Purpose of this Report 2
1.3. Literature Review of Overseas Experience 2
1.4. Local Stakeholder Engagement 3
1.5. Broad Guiding Principles 3
1.6. Key Parameters for CDC in Hong Kong 4
1.7. Structure of this Report 6
2. Project Management Plan (PMP) 7
2.1. Overview of CDC Framework 7
2.1.1. Business and Legal Dimension 10
2.1.1.1. Development objectives 10
2.1.1.2. Legislative framework 14
2.1.1.3 Institutional governance 14
2.1.2. Data Dimension 27
2.1.2.1 Scope of Data 27
2.1.2.2 Database Design for the Two Priority Areas 28
2.1.2.3 Data Alignment Plan 53
2.1.2.4 Data Management and Governance 60
2.1.3. Technology Dimension 72
2.1.3.1 IT Framework 72
2.1.3.2 Data Exchange with Third Parties 77
2.1.3.3 Estimated Size of CDC Database 84
2.1.4 Implementation Dimension 89
2.1.4.1 Implementation Roadmap 89
2.1.4.2 Potential Pilot Projects on "Children with Risk of Abuse and Neglect" and "Children with SEN" 94
2.1.4.3 Project governance 101
2.2. Qualitative Benefits for CDC's Implementation 116
Glossary 122

Appendix A - Summary of Views Gathered from Stakeholder Engagement..... 125
Appendix B – Summary of Findings from Review of Overseas Practices 131
Appendix C – List of Indicative Areas for Analysis 135
Appendix D – Examples of Data Raised by Stakeholders 139
Appendix E - Anticipated Relations and Contributions of CDC Platform Upon Data Alignment 140
Appendix F - Process Diagram for Three Stages of Data Alignment..... 143
Appendix G - List of Proposed Software for Data Analysis 146
Appendix H – Summary of Current State of Relevant Databases of External Parties. 147
Appendix I – Approach and Assumptions for Sizing Estimation..... 152
Appendix J – Case Illustration – Data Linkage Study on Child Abuse..... 156

List of Tables and Figures

Figure 1 Overall CDC Development Framework 7
Figure 2 Overview of Key Distinctive Features for CDC Design 9
Figure 3 Proposed Institutional Setup of CDC 15
Figure 4 Indicative Structure of CDC Project Office 20
Figure 5 Anticipated Relations and Contributions of CDC Platform Upon Data Alignment 48
Figure 6 GSBPM Processes 60
Figure 7 High-level Diagram of the DDI 62
Figure 8 Approval Process for Data De-identification and Data Processing 65
Figure 9 Proposed Technical Architecture under the IT Framework for Foundational Mode 72
Figure 10 Proposed Technical Architecture under the IT Framework for Enhancement Mode 74
Figure 11 Software Components Supporting the Foundational and Enhancement Modes 75
Figure 12 High-level Physical Design Diagram 77
Figure 13 Indicative Reporting and Monitoring Line for CDC 101
Figure 14 Risk Mitigation Prioritisation Matrix 103

Table 1 Developmental Stages and Age Range of General Children 11
Table 2 Potential List of Research Questions and Corresponding Databases 12
Table 3 Proposed Bodies for Oversight, Governance, and Management of CDC 18
Table 4 Proposed Structure for CDC Project Office 23
Table 5 Information on Relevant Databases 28
Table 6 Categorisation and Summary of Variables 32
Table 7 Examination of Data Structure's Consistency for Demographic Data Recorded by Multiple B/Ds' Systems 36
Table 8 Proposed Data Structure for Age, Date of Birth, Ethnicity, Gender, and District of Residence 37
Table 9 Examination of Data Structure's Consistency for Four Types of Abuse and Neglect 40
Table 10 Proposed Data Structure for Neglect, Physical Abuse, Psychological Abuse, and Sexual Abuse 40
Table 11 Examination of Data Structure's Consistency for Hospitalization 40
Table 12 Proposed Data Structure for Hospitalization 41
Table 13 Examination of Data Structure's Consistency for 12 SEN Types 42
Table 14 Proposed Data Structure for 12 SEN Types 43
Table 15 Examination of Data Structure's Consistency for Six Other Child-Related data 44
Table 16 Proposed Data Structure for Six Other Child-related Data 45
Table 17 Number of Records in B/Ds System 51
Table 18 List of Potential Data Items to be Aligned 53
Table 19 Three Stages of Data Alignment 57
Table 20 Mapping between the GSBPM and DDI 63

Table 21 Incremental Approach for Data Alignment....69
Table 22 Overview of Relevant Databases....79
Table 23 Four stages of data exchange with third parties....84
Table 24 Estimated Sizing of CDC under the Foundational Mode....85
Table 25 Projected Storage Requirements for CDC under Foundational Mode....85
Table 26 Server Components and Shared Services under Foundational Mode....86
Table 27 Estimated Sizing of CDC under the Enhancement Mode....87
Table 28 Projected Storage Requirements for CDC under Enhancement Mode....87
Table 29 Server Components and Shared Services under Foundational Mode....87
Table 30 Key Tasks to Be Completed by Phases....91
Table 31 Illustrative RQ for Children with Risk of Abuse and Neglect....94
Table 32 Illustrative RQ for Children with SEN....97
Table 33 Implementation Timeline and Key Tasks for the Illustrations....100
Table 34 Indicative Reporting and Monitoring Mechanism for CDC....101
Table 35 Escalation Path for CDC Implementation Issues....104
Table 36 Potential Risks and Mitigation Strategies of CDC Development....105
Table 37 Indicative Stakeholder Engagement Plan....110
Table 38 Indicative Change Management Plan....112
Table 39 Project Constraints of CDC....113
Table 40 Critical Success Factors of CDC....114
Table 41 Potential Direct Benefits of CDC for Key Users....116

1. Introduction

1.1. Project Overview

In response to the 2017 Policy Address, the Government of the Hong Kong Special Administrative Region (Government) has established the Commission on Children (CoC) on 1 June 2018 to amalgamate the efforts made by relevant Bureaux and Departments (B/Ds) and child concern groups. The vision of CoC is to ensure that Hong Kong is a place where all children's rights, interests and well-being are respected and safeguarded, their voices are heard and where all children enjoy healthy and happy growth and optimal development so as to achieve their fullest potential[1]. The target group will cover children below 18 of age, with a focus on children aged 14 or below.

Prior to its establishment, a Preparatory Committee has been set up by the Government with the objective of gathering views on arrangements for establishing CoC and the priority issues to be handled by CoC. The Preparatory Committee has gathered a significant portion of views that suggested a need to explore the feasibility and implementation framework of developing a Central Databank on Children (CDC). This will enable B/Ds and/or relevant non-government stakeholders to collect and share useful data on children with a view to:

- Collect information of specific categories of children who bear high risk characteristics that have been widely supported by evidence-based research and would need additional and tailored support;
- Establish an effective on-going mechanism of consolidating and integrating data from different stakeholders; and
- Derive useful anonymous data analyses and proactive insights on relevant social issues to formulate policy directions, strategies, recommendations, priorities on child-related issues and improvement of children's services.

Against this backdrop, the Labour and Welfare Bureau (LWB) of the Government has engaged PricewaterhouseCoopers Advisory Services Limited (PwC) to conduct a consultancy study for developing a CDC (the "Study" hereafter). The scope of work for this Study is set out as follows:

- To review and compile a list of existing local databases on children;
- To review overseas experience in developing various types of CDC;
- To formulate a plan for public engagement and provide support in conducting the public engagement exercise;

[1] Terms of Reference of CoC <https://www.coc.gov.hk/en/member.html>

- To process the collected views, including inputs, validation, tabulation and transcription collected from the questionnaire and online surveys, interviews, focus group discussions, forums and/ or other appropriate means;
- To analyse the views collected and provide expert comments and advice on the findings and other related issues;
- To draw up commonly agreed guiding principles for developing CDC;
- To conduct research and assessment on legal, privacy, security, IT (Information Technology) and other key issues involved in the successful development and effective operation of CDC;
- To make recommendations on what type of databank Hong Kong needs and how to develop and maintain CDC; and
- To propose a Programme Management Plan for addressing key and critical issues such as:
    - Projects and tasks to be delivered within specific time frames as well as the interrelationships among these projects/ tasks and their relative dependencies;
    - Mechanism on project governance, risk management, stakeholder management, change management and publicity for development of CDC;
    - Solutions on how CDC interfaces/ integrates with existing and planned government IT infrastructure and systems.

1.2. Purpose of this Report

This Final Report (Report) is the fifth deliverable of the Study. The purpose of this Report is to (i) report validations findings from follow-up meetings with key stakeholders, (ii) document a set of refined Guiding Principles for CDC's development, (iii) formulate a Project Management Plan (PMP) and (iv) present the qualitative benefits of developing a CDC in Hong Kong.

Views gathered from the host of engagement activities conducted in Phase 3 of the Study in relation to various developmental components of CDC are summarised in **Appendix A** for reference.

1.3. Literature Review of Overseas Experience

Findings from the review of five overseas CDC projects (two in the United Kingdom (UK), two in Australia and one in Canada) conducted at early stage of the Study suggested a CDC designed primarily for "trend monitoring" on the one end of the spectrum or for "case tracking" on the other end will have different implications when considering the type of data collection, which will in turn incur very different levels of data privacy concerns. For "trend monitoring", collection of aggregate and anonymised data by leveraging existing operating databases should suffice; whereas "case tracking" will require personal data which are identifiable and traceable. (See **Appendix B** for summary of findings from review of overseas practices).

1.4. Local Stakeholder Engagement

Views of local stakeholders were collated through the following means.

- 20 Interviews with Government Bureaux/Departments (B/Ds), non-governmental organisations (NGOs) and the academia;
- 16 Focus Groups with 64 organisations from the social welfare sector, family groups, schools/educators as well as the social science and healthcare fields;
- 5 Engagement Sessions with over 100 attendees including children, ethnic minorities, parents of children (including parents of children with special educational needs (SEN)) and the general public; and
- Over 1 000 respondents for the Survey.

Details are set out at **Appendix A**.

1.5. Broad Guiding Principles

Taking into account the literature review of overseas experience and the views collated from the local stakeholder engagement exercise, the adoption of the following guiding principles is recommended:

a) ***Purpose-specific/policy-driven:*** Overseas experience suggested that the setting of clear objectives is a key first step in the establishment of any CDC. This view is also supported by the stakeholders engaged for this Study, with many highlighting the importance of defining the objectives and priority areas for CDC development. CDC development should therefore be purpose-specific with clearly defined objectives. Due consideration should be given to its potential in contributing to the attainment of the strategic goals of CoC.

b) ***Privacy protected and secured***: A strong consensus emerging from the stakeholder engagement exercise is that privacy and data security are critical matters to be addressed in order to gain support for and public confidence in the development of a CDC. Having regard to the observed overseas experience and to address public concerns, data privacy and system security of CDC should be accorded the utmost importance when considering the type of databank and data governance mechanism for CDC development. CDC development should also take into consideration stakeholders' expectation as well as the latest development in legislation (e.g. Personal Data (Privacy) Ordinance (Cap. 486)) and measures in relation to data protection.

c) ***User-centric:*** Both findings from the review of overseas experience and views collated through the stakeholder engagement exercise highlighted the importance for CDC to adopt a user-centric design with features that best meet the needs of potential users. Where practicable, it should incorporate features that are valued by users (e.g. data discovery, data visualisation and upload/download functions etc.). In meeting the needs of key users with keen interest in children's well-being in Hong Kong, care should be taken to develop the IT framework and technical architecture of CDC so that collaboration with third parties could be realised.

d) ***Collaborative, Transparent and Consultative:*** Stakeholders generally viewed improving children's well-being as an endeavour that requires cross-sectoral and multi-disciplinary effort. To build public trust for CDC development,

many highlighted the need for setting up a transparent consultative process during CDC development. This view is also in line with observed overseas practices whereby cross-sectoral collaboration and public consultation tend to be adopted when developing CDC. It is therefore suggested that CDC in Hong Kong should be developed as a multi-disciplinary and collaborative effort that involves key stakeholder groups working in sectors/fields related to children's well-being. The process of developing CDC should be transparent to the public, with the provision of a consultative channel for seeking key stakeholders' views on key areas related to its implementation.

e) ***Scalable and incremental:*** CDC development is a complex project that cuts across a wide spectrum of children-related policies/programmes under different B/Ds. As such, it will likely span across a number of years as demonstrated by observed overseas experience. In general, overseas experience suggests that it would take approximately three to six years for the deployment and development of CDC (with variation in timeline depending on factors such as the development objectives of the individual CDC or the types of data collected, etc.). It is therefore suggested that CDC should be developed in phases and in an incremental, scalable manner, including tasks and projects that can demonstrate its benefits.

f) ***Relevant and adaptable:*** As supported by findings from stakeholder engagement and overseas experience, CDC development is a sustained project rather than a time-limited exercise. There is a need to gather feedback on an ongoing basis. Development of CDC should be relevant, timely and adaptable to the evolving needs of children over time, hence the need for regular reviews to assess the development objectives and/or adjust the scope/types of data to be collected.

g) ***Net value to target beneficiaries:*** Views collected through the stakeholder engagement exercise highlighted the importance of appropriately weighing the potential benefits and costs for CDC development, particularly when determining the development objectives and appropriate types of data for collection. It is therefore suggested that the "net value to target beneficiaries" should be adopted as a key consideration in the cost-and-benefit analysis for CDC development in Hong Kong.

1.6. Key Parameters for CDC in Hong Kong

Having applied the broad guiding principles in designing the CDC in Hong Kong, the following parameters are recommended:

a) ***Development Objectives****:* The primary development objective of "trend monitoring" on the basis of non-identifiable data for the general children population should be adopted, while "prevention & early intervention" should only be adopted for specific segment(s) of the children population. In view of the concerns over data privacy collated at the stakeholder engagement exercise and the observations in **Section 1.3**, "case tracking" for the general children population should not be pursued.

b) ***Data Governance****:* While only non-identifiable data will be collected, a special procedure should be developed to safeguard personal data of the children concerned. Apart from compliance with the Personal Data (Privacy) Ordinance (Cap. 486), additional safeguards should also be put in place. This includes separate

data storage and management (see elaboration in point (c) below) and institutional and operational safeguards below :

- In the event that personal data of children will need to be de-identified and linked in data linkage projects for the purpose of realising the development objective of "prevention & early intervention", a Data Ethics and Privacy Panel could be considered for approving the purpose and usage of data collection, with amendments to be made to the Personal Information Collection Statement (PICS);
- Researchers responsible for conducting data linkage projects shall not have access to personal and identifiable data at any given point of time and measures should be put in place to ensure that data subjects will not become identifiable through reverse engineering of de-identified data; and
- There should be segregation of functions whereby different parties should be responsible for de-identifying, linking and using children data for research purpose.

An illustration of the key procedures for approving and conducting data de-identification and linkages encompassing the abovementioned safeguards is shown at **Figure 8**. (See **Section 2.1.1.3** for key institutional setups responsible for the key procedures).

c) ***Data Storage and Management****:* To support data linkage as mentioned above, a federated model with a two-tier database structure should be adopted. At the "macro-level", a central databank with non-identifiable data from Government bureaux and/or departments (B/Ds) will be kept for the purpose of "trend monitoring"; and at the "micro-level", the relevant B/Ds will retain data with identifiable data in their respective databanks for the purposes of "causal analysis and risk identification" (See elaboration in **Section 2.1.3.1**). Overall, established practice should be adopted with regard to data management and governance of CDC (e.g. the UNECE/Eurostat/OECD Generic Statistical Business Process Model (GSBPM)) to meet the specific needs of CDC operations and associated data lifecycle.

d) ***Incremental approach****:* With reference to the development timeframe for CDC projects in overseas countries which took 10 to 15 years (see **Appendix B**), it is recommended that the development of a CDC in Hong Kong should take about ten years in two phases. This indicative development timeframe has also taken into account the lead time required to identify appropriate set of "key data" and for B/Ds to adhere to any agreed practices/guidelines for data alignment to ensure that there would be minimal disruption to the operations of B/Ds while ensuring that new data alignment requirements will be duly considered during new development and/or revamp of IT systems. Taking into account the views expressed by

stakeholders in the engagement exercise in six priority areas[2], the WG has selected "Children with Risk of Abuse & Neglect" and "Children with SEN" as the priority areas, on the consideration that: (i) the target segment can be clearly defined, and data could be relatively easy to collect and retrieve for meaningful analysis and policy formulation; (ii) relevant data of the target segment are currently captured by multiple Government departments/parties; and (iii) the potential impacts would be significant and measurable. The execution of pilot projects is also anticipated to be key in providing recommendations on data to be prioritised for alignment (See the analyses of the risks and opportunities of database design for the two priority areas in **Sections 2.1.4.3.2 and 2.1.1.1.3** respectively). Key tasks to be conducted and the responsible parties involved under the incremental approach are set out in **Table 22**.

1.7. Structure of this Report

This Report consists of the following sections:

Sections	Descriptions
Section 1	Introduction of the Report, including Project Overview, Purpose, Literature Review of Overseas Experience, Local Stakeholder Engagement, Broad Guiding Principles, Key Parameters for CDC in Hong Kong and Structure of this Report.
Section 2	A PMP with elaboration on the feasible implementation framework in the short and medium-to-long run as well as the potential benefits arising from its development.

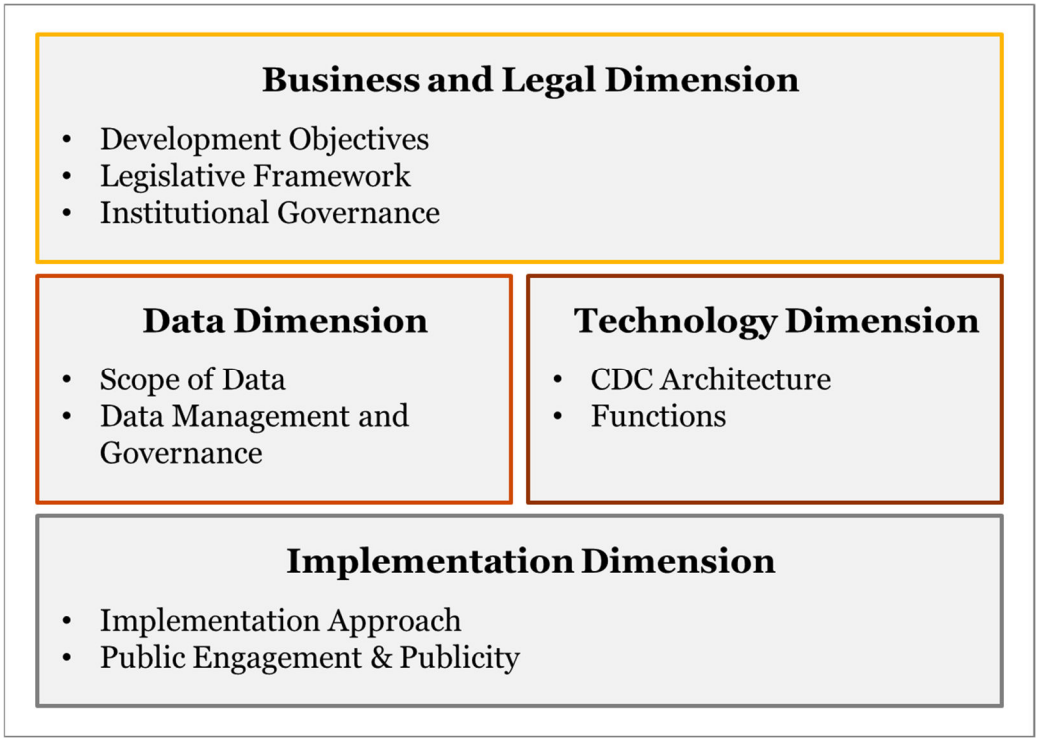
[2] The six priority areas include "Children with Risk of Abuse & Neglect", "Children with Special Educational Needs", "Chronic Health Conditions", "Living in Poverty", "Special Family Background" and "Ethnic Minorities".

2. Project Management Plan (PMP)

2.1. Overview of CDC Framework

The overall CDC framework consists of four board dimensions, i.e. Business and Legal Dimension, Data Dimension, Technology Dimension and Implementation Dimension, as shown in **Figure 1** below. Each dimension will be discussed in detail in the following sections.

Figure 1 Overall CDC Development Framework



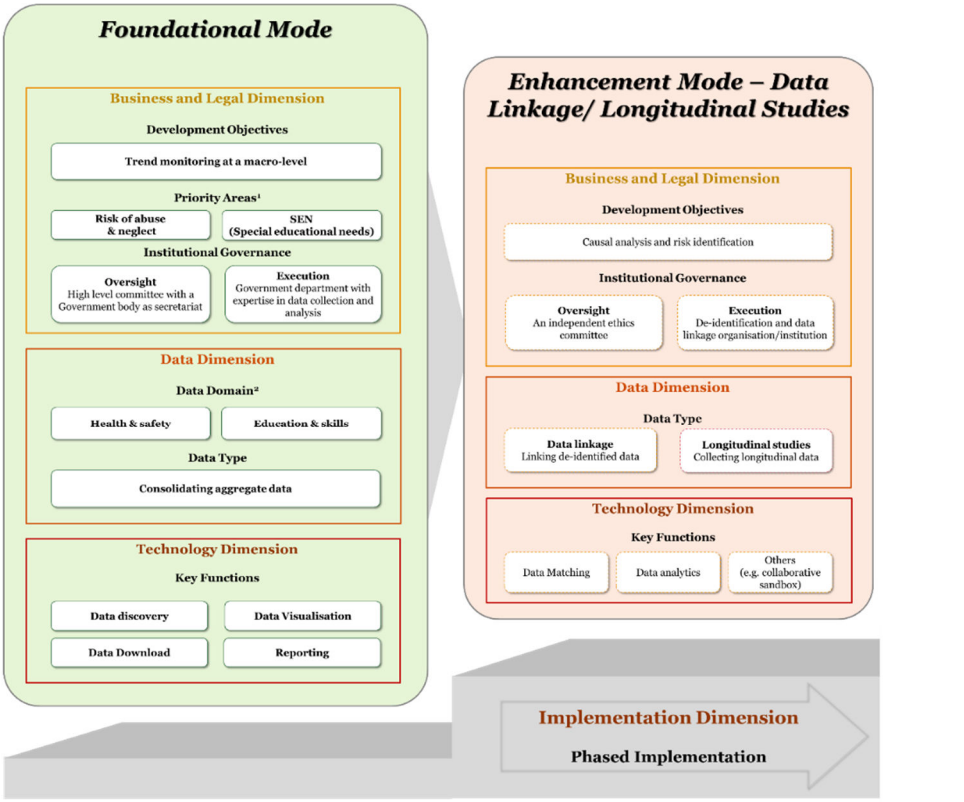
Brief description of each dimension is as follows:

- **Business and Legal Dimension** – This dimension is foundational for CDC development, particularly for steering the direction of CDC development, governing its overall execution; and creating an enabling environment for its implementation. The three key components covered under this dimension are development objectives (including priority areas for data linkages to be elaborated in **Section 2.1.1.1**), legislative framework and institutional governance. All three components have significant implications on the Data, Technology and Implementation Dimensions of CDC development. While development objectives will guide the overall direction and uses of CDC, a suitable legislative framework will be required to ensure secured and appropriate collection, sharing and usage of data. Apart from considering the nature, functions and capability of the executive body for operating CDC, an appropriate governance body will also need to be considered for steering, monitoring and reviewing the progress of CDC development.

- **Data Dimension** – This dimension covers the scope of data (i.e. data domain and data type) to be considered for CDC development as well as the associated data management and governance mechanism for handling the required scope of data. In particular, the scope of data will be determined upon taking into account the development objectives as defined under the Business and Legal Dimension. Once the scope of data has been agreed upon, an appropriate data management and governance mechanism shall be established. The scope of data would also inform the database design (i.e. categories of data, key tasks, contributions and limitation) for the two priority areas and the data alignment plan.
- **Technology Dimension** – This dimension covers the potential technology architecture and functions to be considered for CDC development. Components under this dimension will be defined upon taking into account the components under the Business and Legal Dimension as well as the Data Dimension. In addition to the IT Framework for the Foundational Mode and Enhancement Mode, the function of data exchange with third parties and the estimated size of CDC database will also be addressed under this Dimension.
- **Implementation Dimension** – This dimension covers the components of implementation approach as well as public engagement and publicity for executing CDC development. Once components under the first three dimensions are defined, components under this dimension will set out a practical and feasible implementation approach, with the aim of raising public awareness and acceptability of CDC.

Based on the framework above, two modes of CDC development have been derived to be implemented in different phases across the four dimensions as illustrated in **Figure 2** .

Figure 2 Overview of Key Distinctive Features for CDC Design



(1) It is to note that this does not mean that trend monitoring should only cover the two segments. Rather, a specific focus may be put on the two groups while conducting monitoring.
(2) CDC may collect existing aggregate data pertaining to any data domain at the outset of its implementation, but with a particular emphasis on "health and safety" and "education and skills".

Note 1: For the Foundational Mode, attention shall be paid to the consolidation of relevant aggregate data across B/Ds with emphasis on macro trend monitoring of children in general and the two priority areas of "Children with Risk of Abuse and Neglect" and "Children with SEN".

Note 2: CDC may collect existing aggregate data pertaining to any data domain at the outset of its implementation, but with a particular emphasis on "healthy and safety" and "education and skills".

Note 3: While the Foundational Mode focuses more on data sharing among B/Ds, the Enhancement Mode allows for data sharing with external parties with an approval mechanism in place.

Brief description of each mode is as follows:

- **The Foundational Mode** covers the development objective of "trend monitoring at a macro-level". It will serve as a starting point for CDC development. Component options included in the Foundational Mode are considered more practical, feasible and publicly acceptable to be implemented in the near term.
- **The Enhancement Mode** covers the development objective of "causal analysis and risk identification" beyond the Foundational Mode. Component options considered under this mode will facilitate the execution of either data linkage projects or longitudinal studies, which are considered more practical and feasible to be implemented in the longer term.

2.1.1. Business and Legal Dimension

2.1.1.1. Development objectives

2.1.1.1.1 Foundational Mode

Overall, CDC shall fundamentally be developed with the vision and mission of CoC in mind. In other words, the design of CDC should be able to support CoC's vision of providing a safe and flourishing environment for the healthy and happy growth as well as optimal development of children to enable them to achieve their fullest potential possible. Taking into account stakeholders' views on anticipated benefits from CDC development, CDC is also envisaged to ultimately bring about key benefits such as providing a comprehensive view of children's well-being, facilitating cross-sector collaboration, improving children's service planning and raising public awareness of specific child-related issues.

It is nonetheless noted that CDC will likely enable the realisation of these benefits to a different extent depending on the development objective that is agreed upon.

In alignment with the overall CoC's vision and the expected benefits, the key development objective of the Foundational Mode of CDC is therefore proposed to be "trend-monitoring at a macro-level". Given the use of aggregate and anonymised data, this development objective is likely to face relatively low privacy concern as well as high public acceptance and low controversy in implementation when compared with other development objectives.

Particularly, there should be consideration on the specific topics for meaningful macro-level analysis which could shed light on key well-being issues of children; and preferably, offer insights for systemic policy formulation and service planning. There could be a dual track approach to facilitate the establishment of a set of indicators for monitoring developmental progress of children in Hong Kong. As a starting point, CDC could consider monitoring key trends of general children by developmental stage and of specific segments of children regarding the two priority areas of "children with risk of abuse and neglect" and "children with Special Educational Needs (SEN)", which will be elaborated in **Section 2.1.1.1.3**. As regards macro-level analysis, areas of analysis in which stakeholders engaged have expressed interests are summarised and tabulated at **Appendix C.**

In terms of monitoring key macro-trends of general children by developmental stage, it could be done by analysing aggregate data of general children by their physical development which could be divided by age range, such as pre-schoolers aged between three to five, children aged between six to eleven, and teenager aged between twelve to seventeen. The developmental stages and age range of general children are tabulated on the next page.

Table 1 Developmental Stages and Age Range of General Children

Developmental stage	Age range
Infants	0-1
Toddlers	1-3
Pre-schoolers	3-5
Children	6-11
Teenager	12-17
Young adults	18

2.1.1.1.2 Enhancement Mode

Beyond the Foundational Mode, CDC could consider enabling the development objective of causal analysis and risk identification as well as enhanced macro-trend monitoring by conducting data linkage projects on specific research topics upon giving due consideration to the latest trends and child-related issues in Hong Kong. To achieve this, linkages will need to be created between different datasets addressing the same topic at the backend prior to reporting of macro-level trends to the public or enabling further causal analysis. Likewise, CDC could consider conducting longitudinal studies to further enable causal analysis and risk identification on specific topics or cohorts of children of concern. The set of conditions which need to be fulfilled prior to the implementation of Enhancement Mode is elaborated in **Section 2.1.4.1.1.**

2.1.1.1.3 Priority Areas for Data Linkages

While "causal analysis and risk identification" will be the key development objective of the Enhancement Mode, it is proposed that pilots on data linkages should be conducted during the early stage of CDC implementation to demonstrate the benefits of such linkages and identify key challenges to be tackled to enable smooth execution of the Enhancement Mode. In particular, concerns over technical matters, such as the need for data alignment on linked data, and privacy concerns over data leakage during the process of de-identification would also need to be addressed.

Following deliberation by the Working Group on Research and Public Engagement[3] under CoC, "**children with risk of abuse and neglect**" and "**children with SEN**" have been selected as the two priority areas for data linkages after taking into consideration the following criteria:

[3] Renamed to the Working Group on Research and Development since 1 January 2023.

- **Ability to clearly define the target segment:** whether the target segment can be clearly defined, and data could be easily collected and retrieved for meaningful analysis and policy formulation;
- **Data being captured by multiple B/Ds:** whether relevant data of the target segment are currently captured by multiple B/Ds; and
- **Magnitude of potential impact:** whether the potential impacts would be significant and measurable.

The CDC Platform upon data alignment can potentially contribute to facilitating research in respect of the pilot projects on "Children with Risk of Abuse and Neglect" and "Children with SEN". To help expedite the process for conducting pilots during CDC implementation in the future, illustrations of applications of data linkage on the two priority areas of "Children with Risk of Abuse and Neglect" and "Children with SEN" have been prepared as part of this Study. **Table 2** below presents the potential list of research questions to be investigated and the corresponding databases that will be of relevance to the set of research questions. (See **Section 2.1.4.2** for more details on the two illustrations).

Table 2 Potential List of Research Questions and Corresponding Databases

Potential List of Research Questions	Database Relevant to the Priority Area
Illustration on Children with Risk of Abuse and Neglect	
• What are the macro-trend and incident rates of reported child abuse cases in Hong Kong? • What are the underlying differences for case reporting across databases from Social Welfare Department (SWD), Hospital Authority (HA) and Hong Kong Police Force (HKPF)? • What are the risk factors of reported child abuse cases? • Are there any geographical differences in distribution of reported child abuse cases?	Department of Health (DH) System for Managing the Assessment of Student Health (SMASH) HKPF Case Management and Investigation System (CMIS) SWD Child Protection Registry (CPR) Central Referral System for Residential Child Care Services (CRSRC) HA Clinical Management System (CMS) / clinical database
Illustration on Children with SEN	
• What is the yearly number of children with SEN in Hong Kong?	DH

Potential List of Research Questions	Database Relevant to the Priority Area
• What are the factors contributing to differences in reporting of children with SEN conditions across various databases? • What are the characteristics of children with SEN that can influence health, safety and/ or education outcomes? • What is the prevalent mode and utilisation of health, educational and social service provision (i.e. across or within District Council districts (DC)/ constituency area (CA)) for children with SEN?	Child Assessment Service Information System (CASIS) Child Health Service System (CHSS) System for Managing the Assessment of Student Health (SMASH) <u>Education Bureau (EDB)</u> Special Education Management Information System (SEMIS) <u>HA</u> Cancer Statistics Query System – Children and Adolescents (CanSQS) Clinical Management System (CMS) / clinical database <u>SWD</u> Child Protection Registry (CPR) Central Referral System for Residential Child Care Services (CRSRC)

Between 27 July and 26 September 2022, codebooks of relevant data variables have been obtained from five B/Ds. The purpose of which is to collect further details on data collected in relevant databases that are specifically relevant to the two priority areas, in order to better consider the key data that would require alignment across B/Ds in developing the illustrations. More details on illustrated database design for the priority areas and the proposed data alignment plan, based on the information collected, are at **Sections 2.1.2.2 and 2.1.2.3** respectively.

It is to note that while the development of these illustrations will serve as a reference point for the potential pilot projects to be conducted at the initial stage of CDC development, the party who is commissioned to undertake the eventual pilot project should be given the flexibility to broaden and/or deepen the research scope and rigor giving due consideration to the evolving needs of children, their specialised knowledge on the priority areas and data availability from relevant B/Ds.

2.1.1.2. Legislative framework

2.1.1.2.1 Foundational Mode

The Personal Data (Privacy) Ordinance (PDPO) will be primarily the legislation that offers principal personal data protection. In the nearer term of CDC development, given that only aggregate and anonymised data will be shared, there is likely to be compliance with PDPO. In addition, the CDC platform shall also adopt commonly recognised standards, including the Interoperability Framework for E-Government, Information Security Management Framework, Government IT Security Policy, Guidelines of OGCIO to ensure compliance with the general privacy policy of the Administration.

In the case for executing the pilots in which de-identified data is involved, additional safeguards that are non-legislative in nature could be considered such as the setting up of appropriate governance arrangement for safeguarding data ethics (See **Section 2.1.1.3** for more details).

There is a need to put an appropriate monitoring mechanism in place to ensure that all data shared with CDC comply with PDPO. Currently, the development of the Ethical Accountability Framework for Hong Kong by the Privacy Commissioner for Personal Data could encourage ethical usage of data.

2.1.1.2.2 Enhancement Mode

Given that de-identified data and longitudinal data could be shared through the Enhancement Mode, additional safeguards would likely need to be put in place in addition to compliance with PDPO.

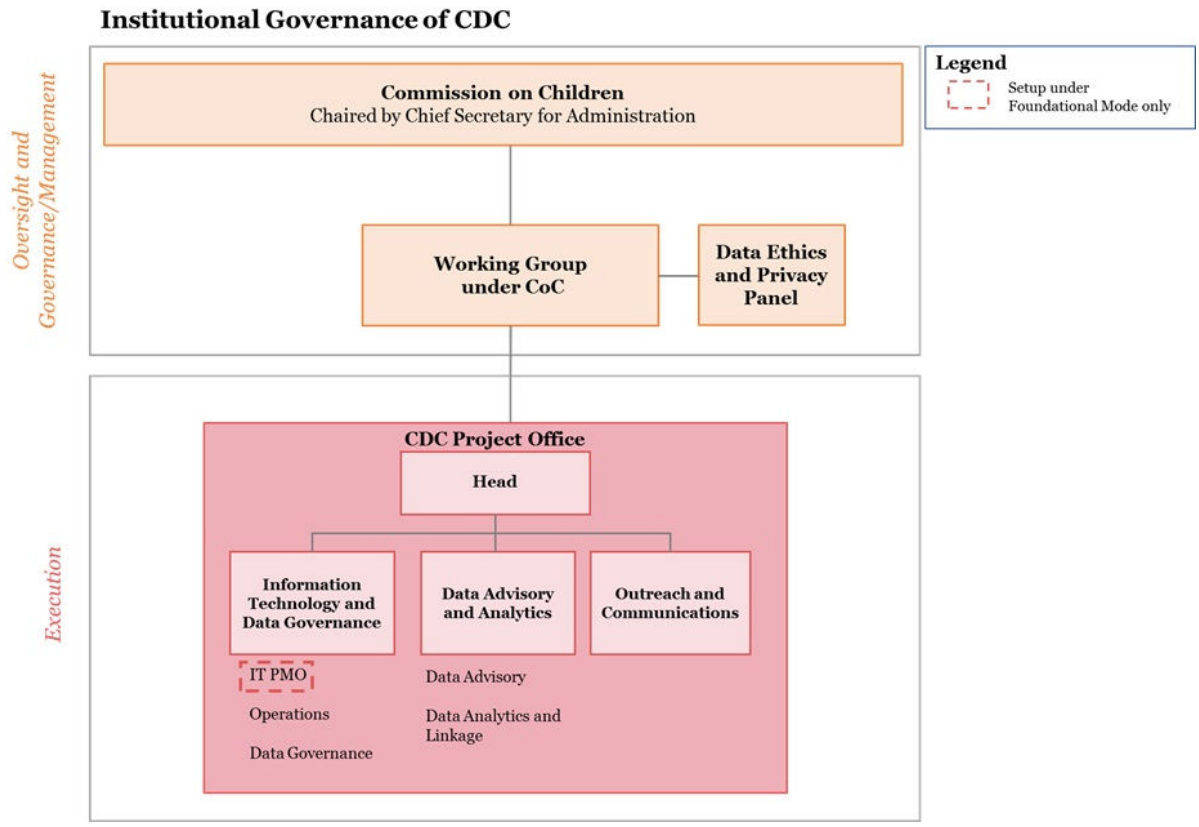
For the sharing of de-identified data, both the Department of Justice (DOJ) and the Privacy Commissioner for Personal Data (PCPD) should be consulted for a more detailed assessment on the privacy and legal implications for sharing of de-identified data, particularly for retrospective sharing. Moving forward, to instil public trust, it is recommended that the Personal Information Collection Statement (PICS) adopted by relevant B/Ds be updated in anticipation of future data sharing opportunities with CDC.

Research ethics would also need to be strengthened, with consideration of ethical implications of commissioned studies (both for data linkage projects and longitudinal studies). Appropriate governance arrangement and guidelines should be put in place to consider the ethical implications of proposed research topics involving children's data. Details on the function of the Data Privacy and Ethics Panel, which is the governance body for research and data ethics, are at **Section 2.1.1.3.3** whereas potential criteria to consider for examining the ethical implications of studies are elaborated at **Section 2.1.2.4.2**.

2.1.1.3 Institutional governance

To deliver the CDC Model, a two-tiered governance structure with clear assignment of roles to entities in undertaking oversight and executive functions of CDC will be established. As shown in **Figure 3**, the establishment will slightly differ under the Foundational and Enhancement Modes in response to the changes on functions performed under the two modes, as explained in the following Section.

Figure 3 Proposed Institutional Setup of CDC



2.1.1.3.1 Foundational Mode

To implement the Foundational Mode of CDC (i.e., with a development objective of "trend monitoring at a macro level"), the two-tiered governance structure as shown in **Figure 3** will be reporting to a relevant working group (WG) under the CoC to ensure that the development of CDC remains in line with the overall policy objective of CoC. Our recommendations on the potential functions, composition as well as skills and expertise to be considered for each of the proposed body are as follows:

Working Group under CoC

It is proposed that CDC be overseen by a relevant WG under CoC, whose functions shall include:

- To provide high-level oversight for CDC development giving due consideration to the strategic priorities of children's well-being of the Government;
- To ensure that the priority areas and scope of data inclusion in CDC development remain relevant and in line with the strategic priorities of children's well-being; and,
- To identify key policy areas to be prioritised for data alignment and analysis, and oversee the formulation of data standards/guidelines including usage of data under the purview of research ethics with reference to local and overseas best practices.

Having regard to the need to perform the above, there could be consideration in incorporating these functions into an existing WG under CoC upon rationalising its current functions and composition before exploring the potential setup of any new WG. Regardless, it is envisioned that the concerned WG shall be chaired by a party outside of the Government to maintain impartiality and independence. Members of the WG shall include non-official members with expertise in wide ranging areas pertinent to the subject including children's well-being (e.g., healthcare, education, family science and social welfare), information and technology (IT), data science, data standards, data privacy and relevant legal subjects.

In addition, representatives from relevant B/Ds shall also be involved as attendees to the WG meetings, including the Census and Statistics Department (C&SD), Office of the Government Chief Information Officer (OGCIO), Office of PCPD, SWD, DH, HA and EDB.

Interim Data Ethics and Privacy Panel

For the purpose of implementing the pilot projects which involve the use of de-identified data, an Interim Data Ethics and Privacy Panel shall be formed to provide relevant support and recommendations to the WG overseeing CDC. As an interim setup, it is suggested that the panel comprises around four to six members involving members from data contributing B/Ds as well as other members with specialist expertise from outside of the Government with expertise or policy experience in the areas of "Children with risk of Neglect and Abuse", "Children with SEN", data privacy and ethics, data analytics, and data anonymisation. In particular, parties with past experience sitting in research ethics review panel would be preferred.

CDC Project Office

A CDC Project Office within the Government will be responsible for implementing the PMP for CDC development under the steer of the WG overseeing CDC. It is proposed that the head of the CDC Project Office shall possess skills/knowledge in realising the potential of CDC and has a keen interest in addressing child-related policy issues, accompanied by a reasonable amount of experience and seniority. It is anticipated that support will be required from technical experts from OGCIO and C&SD. The proposed CDC Project Office shall consist of the following divisions:

- **IT and Data Governance;**
- **Data Advisory and Analytics; and**
- **Outreach and Communications.**

The functions and compositions of the above divisions will be elaborated in the following **Section 2.1.1.3.4.**

2.1.1.3.2 Enhancement Mode

For implementing the Enhancement Mode of CDC (i.e., with a development objective of "causal analysis and risk identification" through the use of data linkage projects and longitudinal studies), it is envisioned that a formalised **Data Ethics and Privacy Panel** will need to be established with a full membership of around nine panel

members[4] in anticipation of the increasing number of research proposals to be examined and approved on a regular basis.

As regards the CDC Project Office, it is anticipated that personnel required for undertaking functions under the IT and Data Governance Division and the Data Advisory and Analytics Division will also be increased to manage the expanded IT operations of CDC as well as to facilitate the execution of a wider number of data linkage projects and longitudinal studies respectively.

2.1.1.3.3 Functions and Composition of the Oversight Bodies

As briefly described earlier, it is envisioned that a **relevant WG under CoC** will take up the role to oversee the CDC.

For the purpose of implementing the pilot projects which involve the use of de-identified data under the Foundational Mode, an Interim Data Ethics and Privacy Panel shall be formed to provide relevant support and recommendations to the WG overseeing CDC. Transitioning to the Enhancement Mode, **a full Data Ethics and Privacy Panel** shall be established to approve data linkage projects and longitudinal studies on ethical grounds on a regular basis in the medium to long term.

The proposed functions and composition of the oversight bodies are elaborated in **Table 3** on the next page.

[4] Reference has been made to the composition of research ethics review panels of local universities and the case of South Australia.

Table 3 Proposed Bodies for Oversight, Governance, and Management of CDC

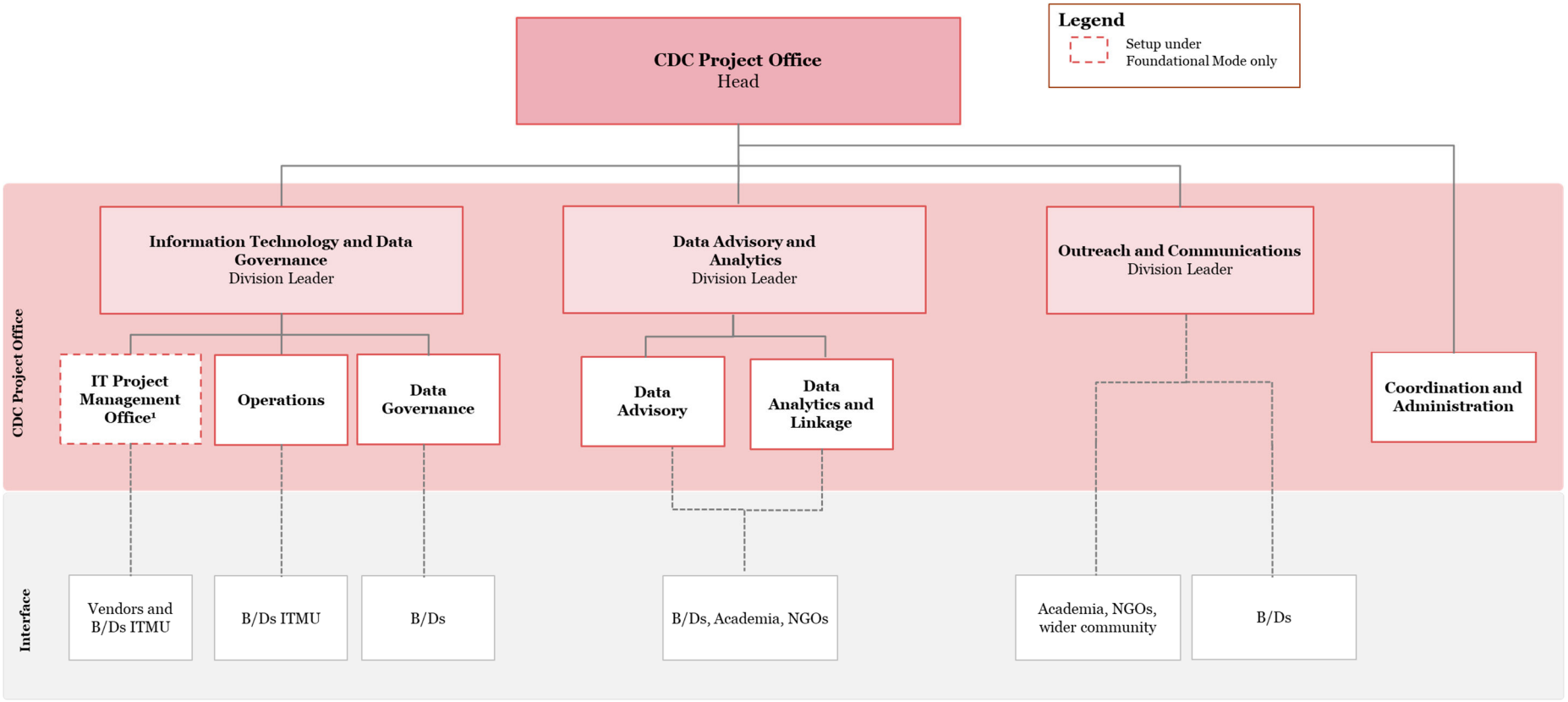
Oversight Bodies	Functions	Potential Composition
A Working Group (WG) under CoC	The WG shall be responsible for providing oversight and steer to CDC on the following: • **To provide oversight to CDC development** in alignment with CoC's strategic priorities of children's well-being; • **To monitor CDC's performance and development progress** regularly to ensure proper and effective performance of CDC; • **To provide directions for research and analysis through CDC** with a view to facilitate evidence-based policymaking for improving children's well-being; • **To ensure the scope of data and priority areas** are in alignment with CoC's strategic priorities for children's well-being; • **To advise on key policy areas** to be prioritised for data alignment and analysis, and the setting of any **data standards/guidelines** for the alignment; and • **To ensure data sharing/ privacy** are in compliance with Personal Data (Privacy) Ordinance (PDPO) and other relevant Government standards. In addition, the WG will also establish CDC's research ethics standards and guidelines with reference to local, overseas and international	The WG shall include representatives with expertise and/or policy experience in the following area: • Children's well-being (e.g. healthcare, education and social welfare); Information and Technology (IT); • Data analytics; • Data standards; and • Personal data privacy. Chaired by a party outside of the Government to maintain impartiality and independence, the WG shall involve non-official members from various sectors. In particular, representatives from the CDC Project Office shall attend meetings of the WG, including: • Head of CDC Project Office; • Head of Information Technology and Data Governance Division; • Head of Data Advisory and Analytics Division; and • Head of Outreach and Communications Division. In addition, representatives from relevant B/Ds shall also be involved as attendees to these committee meetings, including: • SWD; • DH; • HA; • EDB; • C&SD; • OGCIO; • PCPD.

Oversight Bodies	Functions	Potential Composition
	guidelines and practices.	
Data Ethics and Privacy Panel	Under the Foundational Mode, the **Interim Data Ethics and Privacy Panel** will be responsible for the following: • **To examine data privacy and ethical issues** of potential pilot projects; • **To conduct and/or manage ethical and privacy risk assessment** of pilot projects; and • **To provide recommendations on appropriate measures in tackling privacy and ethical issues**	The Interim Panel shall consist of four to six members, including at least one representative with expertise and/or policy experience in each of the following areas: • children with SEN; • risk of child abuse or neglect • data privacy and ethics; • data analytics; and • data anonymisation. Interim Panel may include members from data contributing B/Ds, namely EDB, SWD, DH and HA as well as other members with specialist expertise from outside of the Government. Other members with other expertise could be co-opted for each meeting as necessary. During the interim period, CoC may consider tapping into personnel from a readily available review panel of relevant universities with expertise in SEN and risk of child abuse or neglect.
	Under the Enhancement Mode, a permanent **Data Ethics and Privacy Panel** will continue performing the above functions for future research. In addition, the Panel will be responsible for **providing an annual report** on its activities including setting out the list of projects approved publicly available.	The composition of the permanent Panel shall comprise both Government personnel and external members akin to the Interim Panel, with at least 9 members. Taking reference from the National Statistician's Data Ethics Advisory Committee (NSDEAC) of the UK, it is suggested there shall be no less than 4 external members outside of the Government. A Secretariat shall be established within the Coordination and Administration Unit of the CDC Project Office for the Panel. (Note: According to its Terms of Reference, members of the NSDEAC will include a representative of the National Statistician, at least five independent external members, and no more than four members from the Government.)

2.1.1.3.4 Functions and Composition of CDC Project Office

A CDC Project Office within the Government shall be set up to manage the operations of CDC. The CDC Project Office shall consist of the **Information Technology and Data Governance Division**, the **Data Advisory and Analytics Division**, and the **Outreach and Communications Division.** The below figure presents the proposed structure of the CDC Project Office.

Figure 4 Indicative Structure of CDC Project Office



The proposed functions and composition of the divisions are elaborated below.

Information Technology and Data Governance Division

This division shall be responsible for functions related to IT operations and data governance of CDC. An **Operation Unit** will be responsible for operating and maintaining the CDC platform, whereas a **Data Governance Unit** will be responsible for defining and updating metadata, processing and preparing aggregate data for macro-trend analysis, ensuring regulatory compliance of PDPO and satisfaction of any new data governance requirements (including specified data alignment measures) of the Government, defining access control and creating de-identification keys for the pilot projects on the two priority areas among others.

For the first three years, an interim **IT Project Management Office (PMO)** will be set up to provide dedicated project management for the initial development of CDC.

The division shall be overseen by Division Leader and staffed with personnel with expertise in areas such as project management and procurement, IT solutions, data management, data privacy and data ethics.

Data Advisory and Analytics Division

This division is responsible for data advisory as well as data analytics and linkage functions. It will comprise the **Data Advisory Unit** and **Data Analytics and Linkage Unit**. The former shall provide assistance/guidance to other B/Ds or external parties commissioned to conduct studies on behalf of the Government on potential research questions, research design and data requirement whereas the latter shall coordinate with data owners holding relevant datasets to support data alignment and conversion as well as conduct data linkage for the pilot projects and data linkage projects under the Enhancement Mode.

Overseen by a Division Leader, this division shall be staffed with personnel with expertise in areas such as clinical and social research, statistical analysis, data linkage, predictive modelling and stakeholder coordination.

Outreach and Communications Division

Responsible for outreach as well as communications, this division will be primarily responsible for developing an engagement plan for regular consultation with key external stakeholders (e.g., the academia, non-Governmental organisations (NGOs) and the public), developing and providing training for B/Ds in relation to CDC, and promoting a culture of data-driven insights for child-related policies within the Government.

A Division Leader will oversee this division which shall be staffed with personnel with expertise in areas such as civic dialogue, public relations, marketing, and stakeholder engagement.

Coordination and Administration Unit

Responsible for providing secretariat support to the WG overseeing CDC in relation to matters including the service procurement of the pilot projects, the unit will report directly to the Head of CDC Project Office[5].

The proposed structure, detailed functions, and composition of the CDC Project Office are summarised in **Table 4**.

[5] The Head of the CRC Project Office is expected to be of the highest rank amongst the staff in the Office.

Table 4 Proposed Structure for CDC Project Office

Unit	Functions	Potential Composition
Coordination and Administration	Reporting directly to the Head of CDC Project Office, the Unit shall be responsible for providing secretariat support to the WG overseeing CDC in relation to matters including the service procurement of the pilot projects.	The Unit shall be staffed with personnel with experience and skills in coordination and providing administrative support.
Information Technology (IT) and Data Governance Division will be overseen by a Senior Professional Officer as the Division Leader and is divided into three units: (1) IT Project Management Office, (2) Operations Unit and (3) Data Governance Unit.		
IT Project Management Office (PMO)* ***Interim set-up***	An ad-hoc IT PMO will be set up during the initial stage of implementation (e.g. first three years). Its main responsibility is to: • Purchase necessary procurement items (e.g. software licenses and hardware); and • Manage external contractors on the IT implementation of CDC. • Monitor the IT implementation progress of CDC and conduct System Integration Testing (SIT) and User Acceptance Testing (UAT)	The IT PMO shall be staffed with personnel with expertise in project management and procurement, involving the following roles: • Project Manager; • Analyst *; and • Technical Officer* *Staff will take up role in operations unit in the full CDC Project Office.
Operations Unit	The Unit shall be responsible for the following: • Implement data-related and general functions (e.g. data discovery, visualisation, download and reporting, operating and maintaining online portal for application and feedback); • Implement security-related measures under the advice of the WG overseeing CDC (e.g. access control and sandboxing, audit and logging); • Maintain backup and recovery of the system;	The Unit shall be staffed with personnel with expertise in IT solutions, data management, involving the following roles: • Project Manager; • Analyst; • Technical Officer; and • Computer Operator.

Unit	Functions	Potential Composition
	• Provide administrative support and on premise/ on site management; and • Provide technology support for users (e.g. help desk).	
Data Governance Unit	The Unit shall be responsible for the following: • Define and update metadata and data standards; • Process and prepare aggregate data through coordinating with key B/Ds holding the necessary data for macro-trend analysis or pilot projects; • Define access control; • Ensure regulatory compliance (e.g. PDPO) and satisfaction of data governance requirements (e.g. including personal data impact (PIA) assessment, storage and retention requirements) that are specified in new and emerging regulations; and • Create a de-identification key (or conversion algorithm for facilitating the de-identification of data for linkage and/ or hold the master key the algorithm.	The Unit shall involve personnel with the following skills and expertise: • De-identification key (or conversion algorithm); and • Data governance and knowledge of current trend of data governance requirements such as Personal Data (Privacy) Ordinance (PDPO). The Unit shall involve the following roles: • Data Ethics and Protection Officer • Project Manager; • Analyst; and • Statistician.
Data Advisory and Analytics Division will be overseen by a Senior Professional Officer as the Division Leader and is divided into two units: (1) Data Advisory Unit and (2) Data Analytics and Linkage Unit.		
Data Advisory Unit	The Unit shall be responsible for the following: • Assist and guide B/Ds or external parties commissioned by the Government to conduct studies on appropriate research questions and hypothesis, research methods, data	The Unit shall be staffed with personnel with expertise in the following: • Children's well-being, with a particular focus on child abuse/ neglect and SEN

Unit	Functions	Potential Composition
	requirements and interpretation of children's data; • Liaise between research parties and data-contributing B/Ds • Act as the project manager for the procurement of longitudinal studies conducted by independent tertiary institutions.	• Research matters; and • Experience coordinating with B/Ds. The Unit shall involve the following role: • Researcher.
Data Analytics and Linkage Unit	The Unit will be responsible for: • Coordinate and negotiate with respective data owners holding relevant datasets to support data alignment, collection and de-identification; • Conduct data linkage for pilots and data linkage projects; and • Analyse data collected for pilots.	The Unit shall have expertise in the following areas: • Statistical analyses; and • Data linkage; and • Experience coordinating with B/Ds. The unit shall involve the following roles: • Statistician/Statistical Officer; and • Technical Officer.
Outreach and Communications Division will be overseen by a Senior Professional Officer as the Division Leader.		
The Division is responsible for the following: • **Identify relevant stakeholders and appropriate consultation channels;** • **Conduct consultation with both internal and external stakeholders (e.g. Academia, NGOs, public);** • **Develop and provide training for staff from relevant B/Ds in relation to CDC as well as share success stories and lesson**	The Division shall be staffed with personnel with expertise in stakeholder engagement. Other units of CDC Project Office, such as Operations Unit and Data Governance Unit, will be involved in delivering the appropriate training to B/Ds.	

Unit	Functions	Potential Composition
	learnt with regard to data-driven insights for child-related policies on a regular basis.	

2.1.2. Data Dimension

2.1.2.1 Scope of Data

Despite the different focus on the scope of data under the Foundational Mode and Enhancement Mode as elaborated below, in realising the development objectives of CDC, it is anticipated that data from parties beyond the Government could be shared with the CDC provided privacy is secured and the third party possess technical maturity and quality data. Considerations in relation to data exchange is further elaborated at **Section 2.1.3.2**.

2.1.2.1.1 Foundational Mode

Given that the Foundational Mode of CDC will focus on "trend monitoring at a macro level", particular attention shall be paid to the consolidation of relevant aggregate data across B/Ds. Emphasis could be placed on macro trend monitoring of children in general or the two priority areas of "Children with Risk of Abuse and Neglect" and "Children with SEN". Further breakdown of aggregate data by age range, district, ethnicity and developmental stages could also be considered provided that such data breakdown will not reveal the identity of the data subjects. The collection of such aggregate data allows for a strong foundation to generate more in-depth and comprehensive analysis on children in identifying prevailing trends and new research areas. The demonstration of territory-wide children's statistics with the availability of further data breakdown would help raise public awareness on specific topics and pave the way for the use of aggregate and anonymised data in undertaking more advanced analysis in the future. Examples of data to be consolidated as raised by stakeholders are at **Appendix C**.

2.1.2.1.2 Enhancement Mode

Pilots that leverage the use of de-identified data in relation to "Children with Risk of Abuse and Neglect" and "Children with SEN" are also proposed to be conducted in the near term to assess the benefits and impact of such research prior to more wide scale application in the longer term under the Enhancement Mode.

Under the Enhancement Mode, the scope of data could be widened to include linking de-identified data from the databases of different B/Ds and even parties beyond the Government to meet the development objective of "causal analysis and risk identification and facilitate research on specific topics. As a start under the pilots, potential data to be considered for data matching exercise could be the administrative data across different B/Ds (e.g. SWD, HA, DH, EDB and HKPF). Such data could be integrated to identify risks profiles for children with risk of abuse and neglect and understand the current state of children's service provision, such as evaluating the impact of support provided for children with SEN.

After successful execution and demonstration of benefits of data linkage projects with the pilots, the Enhancement Mode could be implemented in a more systematic and structured manner with data linkage projects conducted on a more regular basis. Beyond data linkage projects, collection of data from a designated cohort of children over a period of time (i.e. longitudinal studies) could be considered to further meet the development objective of "causal analysis and risk identification". There will be a need

to consider if the longitudinal study should focus on specific segments or age range of children. In particular, this approach may be useful for examining developmental progress of children over time and investigating issues which take time to manifest its impact on children. Potential data to be considered for longitudinal studies would be the health data from HA and education data from EDB on kindergarten age children. Such data could be collected to monitor and understand early childhood development trends and progress of a cohort of new-born children.

2.1.2.2 Database Design for the Two Priority Areas

2.1.2.2.1 Relevant Databases

As part of developing CDC in the near term, it is proposed that data linkage projects on the two priority areas of "Children with Risk of Neglect and Abuse" and "Children with SEN" can be conducted as pilots. To address the potential list of research questions for the two priority areas of "Children with Risk of Abuse and Neglect" and "Children with SEN" in **Section 2.1.1.1.3,** five B/Ds have provided the codebooks on variables of relevant databases, covering the variable names, variable labels, values, value labels, variable types and definitions if any.[6] Information on the relevant databases are summarised below:

Table 5 Information on Relevant Databases

Department Name	Database Name	Scope
Department of Health (DH)	Child Assessment Service Information System (DH-CASIS)	DH-CASIS which is operated by the Specialised Services Branch – Child Assessment Service (CAS) records the information of children who are under 12 years of age with developmental- behavioural problems or disorders.
	Child Health Service System (DH-CHSS)	DH-CHSS which is managed by the Family and Student Health Branch – Family Health Service (FHS) records the information of babies and children from birth to five years.
	System for Managing the Assessment of Student Health (DH-SMASH)	DH-SMASH which is operated by the Family and Student Health Branch — Student Health Service (SHS) records the information of primary and secondary school students.

[6] "Variable name" refers to the name assigned to each variable; "Variable label" refers to a brief description of each variable; "Values" refers to the actual coded values in the data for each variable; "Value labels" refers to the textual descriptions of the coded values; "Variable types" refers to the types of each variable, i.e. numeric, string; and "Definitions" refers to definitions or international classifications of the variable if any.

Education Bureau (EDB)	Kindergarten Education Scheme System (EDB-KGESS)	EDB-KGESS which is managed by the School Development and Administration Branch – Kindergarten Education Division (KGED) records the information of children attending the kindergartens joining the Kindergarten Education Scheme.
	Special Education Management Information System (EDB-SEMIS)	EDB-SEMIS which is operated by the Professional Development & Special Education Branch – Special Education Division (SED) records the information of students with special educational needs (SEN) and the services include Educational Audiology Service, School-based Speech Therapy Services, School-based Educational Psychology Service, teacher professional development on catering for students with SEN as well as referral and placement services for aided special schools.
	Student Information Management System (EDB-STIMS)	EDB-STIMS which is managed by the Planning, Infrastructure and School Places Allocation Branch – Education Commission and Planning Division (ECP) records the information of students.
Social Welfare Department (SWD)	Child Protection Registry (SWD-CPR)	SWD-CPR maintained by the SWD collects and compiles statistical information on the children who have been maltreated/might have been maltreated or are currently at risk of maltreatment and the perpetrators/ alleged perpetrators/ potential perpetrators for the purpose of ascertaining the magnitude of the problem, including identification of the general profile and characteristics of child maltreatment.
	Central Referral System for Residential Child Care Services (SWD-CRSRC)	The SWD-CRSRC is maintained by SWD to record the demographic data of children who have been waitlisted/are waitlisting for residential childcare services, including personal particulars, family and housing type, schooling and health condition, etc.
Hospital Authority (HA)	Cancer Statistics Query System – Children and Adolescents (HA-CanSQS).	Hong Kong Cancer Registry (HKCaR) is a population-based registry committed to collecting and conducting analyses on data from all cancer cases in Hong Kong and is responsible for managing the database of HA-CanSQS – Children and Adolescents.

	Clinical Management System (CMS) / clinical database	It records the clinical data of the patient from all public hospitals including inpatient, outpatient and emergency department admissions information.
Hong Kong Police Force (HKPF)	Case Management and Investigation System (HKPF-CMIS)	HKPF-CMIS is managed by the Information Systems Wing, HKPF. It records case information including but not limited to those related to family violence, domestic violence, sexual violence, child abuse, elder abuse and juvenile crime.

2.1.2.2.2 Categorisation of Key Data

The variables of the codebooks from the databases are classified into three main categories, including (1) demographic data; (2) subject-specific data; and (3) other child-related data. These categories could be further classified as follows:

- demographic data can be classified as: (1a) demographic data which are recorded by multiple B/Ds' systems; and (1b) demographic data which are recorded by one B/D's system only;
- subject-specific data can be classified as: (2a) key subject-specific data (i.e., incidence rates of different forms of child abuse & neglect) which are recorded and adopted by multiple B/Ds' systems and the approach of which data is defined and collected is methodologically consistent; (2b) other subject-specific data which are recorded by multiple B/Ds' systems but the approach and assessment framework might not be methodologically consistent; and (3b) other subject-specific data which are recorded by one B/D's system only; and

  other child-related data, it can be classified as: (3a) other child-related data related to the policy domains of specific B/Ds' systems but are recorded by multiple B/Ds' systems; and (3b) other child-related data which are recorded by one B/D's system only.

Table 6 on the following page categorises and summarises the variables of the codebooks.

Table 6 Categorisation and Summary of Variables

Data Categories	DH			EDB			HA		HKPF	SWD	
	CASIS	CHSS	SMASH	KGESS	SEMIS	STIMS	CMS / clinical database	CanSQS	CMIS	CPR	CRSRC
(1) Demographic data											
(1a) Demographic data recorded by multiple B/Ds' systems											
Age			✓ (1)							✓ (1)	✓ (1)
Date of birth	✓ (1)	✓ (1)	✓ (1)	✓ (1)	✓ (1)	✓ (1)	✓ (1)		✓ (1)	✓ (1)	✓ (1)
Ethnicity					✓ (1)	✓ (1)				✓ (1)	✓ (1)
Gender	✓ (1)	✓ (1)	✓ (1)	✓ (1)	✓ (1)	✓ (1)	✓ (1)	✓ (1)	✓ (1)	✓ (1)	✓ (1)
District of Residence						✓ (1)	✓ (1)				✓ (1)
(1b) Demographic data recorded by one B/D's system only											
In HK since Birth										✓ (1)	
New arrival										✓ (1)	
Place of birth			✓ (1)								
(2) Subject-specific data – Child Abuse and Neglect											
(2a) Key subject-specific data - Child Abuse and Neglect											
Neglect			✓ (1)				✓ (*)			✓ (8)	
Physical abuse							✓ (*)		✓ (1)	✓ (6)	
Psychological abuse							✓ (*)			✓ (6)	
Sexual abuse			✓ (1)				✓ (*)		✓ (1)	✓ (11)	
(2b) Other subject-specific data recorded by multiple B/Ds' system - Child Abuse and Neglect											
Hospitalisation									✓ (1)		✓ (1)
(2c) Other subject-specific data recorded by one B/D's system only - Child Abuse and Neglect											
Injury information									✓ (3)		
Disclose type										✓ (2)	
Incident location										✓ (13)	
Perpetrator information									✓ (3)	✓ (1)	
Other abuse type			✓ (1)						✓ (4)	✓ (2)	
Maternal use of illicit drugs affecting newborn							✓ (4)				
(2) Subject-specific data – SEN											
(2a) Key subject-specific data – SEN											

Data Categories	DH			EDB			HA		HKPF	SWD	
	CASIS	CHSS	SMASH	KGESS	SEMIS	STIMS	CMS / clinical database	CanSQS	CMIS	CPR	CRSRC
Attention-deficit/hyperactivity disorder (ADHD)	✓ (2)		✓ (2)		✓ (1)		✓ (*)			✓ (1)	
Autism Spectrum Disorder (ASD)	✓ (2)				✓ (1)		✓ (*)			✓ (1)	
Emotional and Behavioural Difficulties (EBD)			✓ (1)				✓ (*)			✓ (2)	
Hearing impairment (HI)	✓ (10)				✓ (2)		✓ (*)			✓ (1)	
Intellectual disability (ID)	✓ (11)		✓ (6)		✓ (2)		✓ (*)			✓ (2)	✓ (1)
Mental illness (MI)	✓ (20)				✓ (1)		✓ (*)			✓ (2)	
Motor impairment (MotorI)	✓ (8)						✓ (*)				
Physical disability (PD)					✓ (1)		✓ (*)			✓ (2)	✓ (1)
Specific Learning Difficulties (SpLD)	✓ (4)				✓ (2)		✓ (*)			✓ (2)	
Speech impairment (SI)	✓ (4)				✓ (1)		✓ (*)			✓ (1)	
Visceral disability and chronic illness (VD&CI)							✓ (*)	✓ (1)		✓ (1)	
Visual impairment (VI)	✓ (4)				✓ (2)		✓ (*)			✓ (1)	✓ (1)
(2c) Other subject-specific data recorded by one B/D's system only – SEN											
Assessment information					✓ (2)						
Borderline intelligence			✓ (1)								
Details at School										✓ (1)	
Diagnosis information	✓ (3)										
Disability information										✓ (2)	
Referral actions			✓ (2)								
(3) Other child-related data											
(3a) Other child-related data recorded by multiple B/Ds' systems											
Health											
Height		✓ (1)	✓ (1)								
Weight		✓ (1)	✓ (1)								
Family											

Data Categories	DH			EDB			HA		HKPF	SWD	
	CASIS	CHSS	SMASH	KGESS	SEMIS	STIMS	CMS / clinical database	CanSQS	CMIS	CPR	CRSRC
Education attainment of father		✓ (1)	✓ (1)								
Education attainment of mother		✓ (1)	✓ (1)								
Housing type			✓ (1)							✓ (1)	✓ (1)
Language at home					✓ (1)	✓ (1)					
(3b) Other child-related data recorded by one B/D's system only											
School											
Details at school			✓ (2)	✓ (5)		✓ (7)					
Education level			✓ (1)								
Kindergarten level				✓ (1)							
Last class attended											✓ (1)
Other child conditions			✓ (8)								✓ (7)
Health											
Category of age at diagnosis								✓ (1)			
Head circumference		✓ (1)									
Number of subjects								✓ (1)			
Year of diagnosis								✓ (1)			
Number of times admitted in hospital in the year							✓ (1)				
Length of stay (in days) in hospital in the year							✓ (1)				
Family											
CSSA										✓ (1)	
Family structure										✓ (2)	✓ (1)
New arrival											✓ (1)
Residential status										✓ (1)	
Date of birth of father		✓ (1)									
Date of birth of mother		✓ (1)									
Occupation of father			✓ (1)								
Occupation of mother			✓ (1)								
Background of mother		✓ (11)									
Other family conditions		✓ (1)								✓ (7)	✓ (1)

✓ refers to the variables that are readily available in the databases

( ) refers to number of variables in the data category

(*) the diagnoses are based on ICD-9-CM,with mapping to 2010 v ICD-10. Review on the diagnosis list by subject experts, which could take up to a few months, is recommended for a more comprehensive diagnosis list for child abuse & neglect, and SEN. The number of variables in the data category are therefore to be determined after the expert review.

2.1.2.2.3 Data Alignment for Demographic Data

Demographic Data Recorded by Multiple B/Ds' Systems

Having consolidated the codebooks of relevant databases, five demographic data (i.e. age, date of birth, district of residence, ethnicity and gender) are found to be recorded by multiple B/Ds' systems. **Table 7** examines the consistency of the data structure across databases including the construct, variable types, values and value labels for these demographic data.

Table 7 Examination of Data Structure's Consistency for Demographic Data Recorded by Multiple B/Ds' Systems

Variables	Data structure (Y – Aligned; N – Not aligned)			
	Construct[7]	Variable types[8] (i.e. string, numeric)	Values[9] (i.e. 1,2,3…)	Value labels[10] (i.e. male, female)
Age	Y	Y	Y	Y
Date of birth	Y	N	N	N
Ethnicity	Y	N	N	N
Gender	Y	N	N	N
District of residence	Y	N	N	N

It is noted that:

- for age, the data structures across systems are aligned;
- for date of birth, the variable types, values and value labels may not be the same;
- for ethnicity, some systems recorded a detailed breakdown, and some did not;
- for gender, some systems used string, and some used numeric; and
- for district of residence, some systems recorded the address, some used string for district coding.

As illustrated in the above, the same types of demographic information are being recorded using different data structures by multiple B/Ds' systems. It is therefore proposed to algin the data structures of these five demographic data. The proposed data structures including variable name, variable type, values and value labels are listed in **Table 8.** As for ethnicity, it is proposed that B/Ds align this variable by adopting the classification of ethnicity used by C&SD.

[7] "Construct" refers to the same aspect or dimensions that the variables are measuring.
[8] "Variable type" refers to the types of each variable, i.e. numeric, string.
[9] "Values" refers to the actual coded values in the data for each variable, i.e. 1, 2, 3.
[10] "Value labels" refers to the textual descriptions of the coded values, i.e. male, female.

Table 8 Proposed Data Structure for Age, Date of Birth, Ethnicity, Gender, and District of Residence

Variables	Proposed data structure			
	Variable name	Variable type	Values	Value labels
Age	Age	Numeric	0-99	0-99
			888	Information not available/ unknown
Date of birth	DOB	String in date format	DDMMYYYY	DDMMYYYY
Ethnicity	Ethnicity	Numeric	1	Chinese
			2	Filipino
			3	Indonesian
			4	Indian
			5	Nepalese
			6	Pakistani
			7	Other South Asian
			8	Thai
			9	Japanese
			10	Korean
			11	Other Asian
			12	White
			13	Black
			14	Others
			888	Information not available/ unknown
Gender	Gender	Numeric	1	Male
			2	Female
			3	Others
			888	Unknown
District of residence	District_of_residence	Numeric	1	Central and Western
			2	Wan Chai
			3	Eastern
			4	Southern
			5	Kowloon City
			6	Kwun Tong
			7	Sham Shui Po
			8	Wong Tai Sin
			9	Yau Tsim Mong
			10	Islands
			11	North
			12	Sai Kung
			13	Sha Tin
			14	Tai Po
			15	Tsuen Wan
			16	Kwai Tsing
			17	Tuen Mun
			18	Yuen Long
			19	Mainland China
			20	Other Places
			888	Information not available/ unknown

Demographic Data Recorded by One B/D's System Only

Three demographic data (i.e. new arrival, place of birth and whether the children is in Hong Kong since birth) are found to be recorded by one B/Ds' system only. Since these data concern only one system, no data alignment is required. These data could be readily available and be provided directly to the consolidated children database.

Key Tasks for Data Alignment

Static data refers to data that do not change after being recorded whereas dynamic data refers to data that may change after being recorded and has to be continually updated. In the case of demographic data, date of birth and ethnicity are static data whereas age and district of residence is a dynamic data. Despite the fact that gender can be subject to change, it is classified as a type of static data in this context given the unlikelihood for gender change to happen before the age of 18.

The key tasks involved in the data alignment are specified as follows:

Step 1 – B/Ds extract the five demographic data (i.e. age, date of birth, ethnicity, gender and district of residence) that are recorded by multiple B/Ds' systems, and other demographic data (i.e. new arrival, place of birth and whether the children is in Hong Kong since birth) that are recorded by one B/Ds' system.

Step 2 – B/Ds compile the extracted demographic data for de-identification.

Step 3 – (Option 1) B/Ds transform the five demographic data (i.e. age, date of birth, ethnicity, gender and district of residence) according to the proposed data structure and submit the de-identified data to the CDC Project Office.

(Option 2) B/Ds submit the de-identified data to the CDC Project Office, and the CDC Project Office transform the five demographic data (i.e. age, date of birth, ethnicity, gender and district of residence) according to the proposed data structure.

While there could be two options on which party to transform the data, Option 1 is preferred as B/Ds have more understanding on the data structure of their systems and some re-coding may involve sensitive data for rare child cases. Taking also into consideration the issue of data sensitivity and confidentiality, B/Ds will be in a better position to conduct the transformation process of the data alignment.

Step 4 – The CDC Project Office checks for consistency of the five demographic data. In the consolidated children database, the new variables of the five demographic data are created.

For static data:
If the demographic data (i.e. date of birth, ethnicity and gender) of a child are matched, the demographic data will be stored in the new variables of the consolidated children database.

If the demographic data of a child are not matched, the CDC Project Office will inform B/Ds to counter check the details.

For instance, if two systems record a child as a girl whereas another system record a child as a boy, the CDC Project Office will inform the B/Ds to counter check the record.

For dynamic data:
If the demographic data (i.e. the district of residence) of a child is matched, the demographic data will be stored in the new variable of the consolidated children database.

If the demographic data of a child is not matched, the CDC Project Office will check the latest records from the B/Ds' systems which will be stored in the new variable of the consolidated children database.

The CDC Project Office checks for the consistency of the demographic data recorded by one B/Ds' system only and these data will be stored in the consolidated children database.

Step 5 – B/Ds counter check the details of those unmatched demographic static data and provide the updated data to the CDC Project Office.

Step 6 – The CDC Project Office updates the data in the consolidated children database. For Step 4 to Step 6 should be repeated until the demographic static data are matched or in some circumstances such as unavailable information and incomplete checking, consensuses have been made across B/Ds on the inconsistent demographic static data to be recorded in the consolidated children database. Discussions or meetings between B/Ds and the CDC Project Office may be required for handling the data inconsistences.

Step 7 – If trend monitoring is required, B/Ds will provide the data including the new entries of the children and updated dynamic data of the children who have already been in the consolidated children database to the CDC Project Office on a regular basis.

2.1.2.2.4 Priority Area for "Children with Risk of Abuse and Neglect"

Key Subject-specific Data

Four types of abuse and neglect (i.e. neglect, physical abuse, psychological abuse and sexual abuse) are found to be recorded by multiple B/Ds' systems upon consolidation of codebooks of relevant databases. **Table 9** presents an examination of the consistency of the data structure across databases for these types of abuse and neglect.

Table 9 Examination of Data Structure's Consistency for Four Types of Abuse and Neglect

Variables	Data structure (Y – Aligned; N – Not aligned)			
	Construct	Variable types (i.e. string, numeric)	Values (i.e. 1,2,3…)	Value labels (i.e. male, female)
Neglect	Y	N	N	N
Physical abuse	Y	N	N	N
Psychological abuse	Y	N	N	N
Sexual abuse	Y	N	N	N

It is found that the variable types, values and value labels of these four types of abuse and neglect are not aligned.

As illustrated above, the same information on the type of abuse and neglect could be recorded using different data structures by the system of multiple B/Ds. It is therefore proposed that the data structures of the four types of abuse and neglect be aligned. The proposed data structures including variable name, variable type, values and value labels are listed in **Table 10**.

Table 10 Proposed Data Structure for Neglect, Physical Abuse, Psychological Abuse, and Sexual Abuse

Variables	Proposed data structure			
	Variable name	Variable type	Values	Value labels
Neglect	Neglect	Numeric	1	Yes
			0	No
Physical abuse	Physical_abuse	Numeric	1	Yes
			0	No
Psychological abuse	Psychological_abuse	Numeric	1	Yes
			0	No
Sexual abuse	Sexual_abuse	Numeric	1	Yes
			0	No

Other Subject-specific Data Recorded by Multiple B/Ds' Systems

The variable denoting whether the child has been admitted in hospital due to the incident of abuse and neglect was found to be recorded by the systems of multiple B/Ds. **Table 11** examines the consistency of the data structure across databases

Table 11 Examination of Data Structure's Consistency for Hospitalization

Variables	Data structure (Y – Aligned; N – Not aligned)			
	Construct	Variable types (i.e. string, numeric)	Values (i.e. 1,2,3…)	Value labels (i.e. male, female)
Hospitalisation	Y	N	N	N

It is found that the variable types, values and value labels are not aligned for this variable.

As illustrated above, the information on whether a child has been admitted is being recorded using different data structures by multiple B/Ds' systems. It is therefore proposed that the data structure of this variable be aligned. The proposed data structure including variable name, variable type, values and value labels is listed in **Table 12**.

Table 12 Proposed Data Structure for Hospitalization

Variables	Proposed data structure			
	Variable name	Variable type	Values	Value labels
Hospitalisation	Hospitalisation	Numeric	1	Yes
			0	No

Other Subject-specific Data Recorded by One B/D's System only

Six categories of other subject-specific data involving 33 variables are found to be recorded by one B/Ds' system only. These data could be readily available and be provided directly to the consolidated children database with no data alignment required.

Key Task for Data Alignment

Both the key data and other subject-specific data mentioned above are dynamic data. The key tasks involved in the data alignment are specified as follows:

Step 1 – B/Ds extract the four key subject-specific data (i.e. neglect, physical abuse, psychological abuse and sexual abuse) that are recorded by multiple B/Ds' systems, other subject-specific data including the one variable (i.e. hospitalisation) that is recorded by multiple B/Ds' systems and other variables that are recorded by one B/Ds' system only.

Step 2 – B/Ds compile the extracted subject-specific data for de-identification.

Step 3 – (Option 1) B/Ds transform the key subject-specific data (i.e. neglect, physical abuse, psychological abuse and sexual abuse) according to the proposed data structure and submit the de-identified data to the CDC Project Office.

(Option 2) B/Ds submit the de-identified data to the CDC Project Office, and the CDC Project Office transform the key subject-specific data (i.e. neglect, physical abuse, psychological abuse and sexual abuse) according to the proposed data structure.

While there could be two options on which party to transform the data, Option 1 is preferred as B/Ds have more understanding on the data structure of their systems and some re-coding may involve sensitive data for rare child cases. Taking also into consideration the issue of data sensitivity and confidentiality, B/Ds will be in a better position to conduct the transformation process of the data alignment.

Step 4 – The CDC Project Office checks for consistency of the key subject-specific data, and these data will be stored in the consolidated children database.

Step 5 – The CDC Project Office derives new variables for the four key subject-specific data according to various research hypotheses where necessary.

Step 6 – If trend monitoring is required, B/Ds will provide the data including the new entries of the children and updated dynamic data of the children who have already been in the consolidated children database to the CDC Project Office on a regular basis.

2.1.2.2.5 Data Alignment for Subject-specific Data on Priority Area for "Children with SEN"

Key Subject-specific Data

12 SEN types are found to be recorded by multiple B/Ds' systems upon consolidation of codebooks of relevant databases. **Table 13** examines the consistency of the data structure across databases.

Table 13 Examination of Data Structure's Consistency for 12 SEN Types

Variables	Data structure (Y – Aligned; N – Not aligned)			
	Construct	Variable types (i.e. string, numeric)	Values (i.e. 1,2,3...)	Value labels (i.e. male, female)
Attention-deficit/hyperactivity disorder (ADHD)	Y	N	N	N
Autism Spectrum Disorder (ASD)	Y	N	N	N
Emotional and Behavioural Difficulties (EBD)	Y	N	N	N
Hearing impairment (HI)	Y	N	N	N
Intellectual disability (ID)	Y	N	N	N
Mental illness (MI)	Y	N	N	N
Motor impairment (MotorI)	Y	N	N	N
Physical disability (PD)	Y	N	N	N
Specific Learning Difficulties (SpLD)	Y	N	N	N
Speech impairment (SI)	Y	N	N	N
Visceral disability and chronic illness (VD&CI)	Y	N	N	N
Visual impairment (VI)	Y	N	N	N

It is found that the variable types, values and value labels of the 12 SEN types are not aligned.

As illustrated in the above, the same information concerning these 12 types of SEN is being recorded using different data structures by multiple B/Ds' systems. It is therefore proposed that the data structures of the 12 SEN types be aligned. The proposed data

structures including variable name, variable type, values and value labels are listed in **Table 14**.

Table 14 Proposed Data Structure for 12 SEN Types

Variables	Proposed data structure			
	Variable name	Variable type	Values	Value labels
Attention-deficit/hyperactivity disorder (ADHD)	ADHD	Numeric	1	Yes
			0	No
Autism Spectrum Disorder (ASD)	ASD	Numeric	1	Yes
			0	No
Emotional and Behavioural Difficulties	Emotional_Behavioural	Numeric	1	Yes
			0	No
Hearing impairment	Hearing_impairment	Numeric	1	Yes
			0	No
Intellectual disability (ID)	Intellectual_disability	Numeric	1	Yes
			0	No
Mental illness	Mental_illness	Numeric	1	Yes
			0	No
Motor impairment	Motor_impairment	Numeric	1	Yes
			0	No
Physical disability	Physical_disability	Numeric	1	Yes
			0	No
Specific Learning Difficulties (SpLD)	SpLD	Numeric	1	Yes
			0	No
Speech impairment	Speech_impairment	Numeric	1	Yes
			0	No
Visceral disability and chronic illness	Visceral_chronic	Numeric	1	Yes
			0	No
Visual impairment	Visual_impairment	Numeric	1	Yes
			0	No

Other Subject-specific Data Recorded by one B/D's System Only

Six categories of other subject-specific data involving 11 variables are found to be recorded by one B/Ds' system only. These data could be readily available and be provided directly to the consolidated children database with no data alignment is required.

Key Tasks for Data Alignment

Both the key data and other subject-specific data are dynamic data. The key tasks involved in the data alignment are specified as follows:

Step 1 – B/Ds extract the four key subject-specific data (i.e. 12 SEN types) that are recorded by multiple B/Ds' systems, and other variables that are recorded by one B/Ds' system only.

Step 2 – B/Ds compile the extracted subject-specific data for de-identification.

Step 3 – (Option 1) B/Ds transform the key subject-specific data (i.e. 12 SEN types) according to the proposed data structure and submit the de-identified data to the CDC Project Office.

(Option 2) B/Ds submit the de-identified data to the CDC Project Office, and the CDC Project Office transform the key subject-specific data (i.e. 12 SEN types) according to the proposed data structure.

While there could be two options on which party to transform the data, Option 1 is preferred as B/Ds have more understanding on the data structure of their systems and some re-coding may involve sensitive data for rare child cases. Taking also into consideration the issue of data sensitivity and confidentiality, B/Ds will be in a better position to conduct the transformation process of the data alignment.

Step 4 – The CDC Project Office checks for consistency of the subject-specific data, and these data will be stored in the consolidated children database.

Step 5 – The CDC Project Office derives new variables for the four key subject-specific data according to various research hypotheses where necessary.

Step 6 – If trend monitoring is required, B/Ds will provide the data including the new entries of the children and updated dynamic data of the children who have already been in the consolidated children database to the CDC Project Office on a regular basis.

2.1.2.2.6 Data Alignment for Other Child-related Data

Other Child-related Data Recorded by Multiple B/Ds' systems

Six other child-related data (i.e. height, weight, education attainment of father, education attainment of mother, housing type and language at home) are found to be recorded by multiple B/Ds' systems upon consolidation of codebooks of relevant databases. **Table 15** examines the consistency of the data structure of these data.

Table 15 Examination of Data Structure's Consistency for Six Other Child-Related data

Variables	Data structure (Y – Aligned; N – Not aligned)			
	Construct	Variable types (i.e. string, numeric)	Values (i.e. 1,2,3…)	Value labels (i.e. male, female)
Height	Y	Y	Y	Y
Weight	Y	Y	Y	Y
Education attainment of father	Y	N	N	N
Education attainment of mother	Y	N	N	N
Housing type	Y	N	N	N
Language at home	Y	Y	Y	Y

It is found that:

- for height and weight, the data structures across systems are aligned;
- for education attainment of father and mother, the variable types, values and value labels are not aligned.;
- for housing type, some systems recorded a detailed breakdown, and some did not; and
- for language at home, the data structure across systems are aligned.

As illustrated above, the same information on the six types of child-related data are being recorded in multiple B/Ds' systems using different data structure. it is therefore proposed to algin the data structures of the six other child-related data. The proposed data structures including variable name, variable type, values and value labels are listed in **Table 16** below.

Table 16 Proposed Data Structure for Six Other Child-related Data

Variables	Proposed data structure			
	Variable name	Variable type	Values	Value labels
Height	Height	Numeric	unit	in cm
			888	Information not available/ unknown
Weight	Weight	Numeric	unit	in kg
			888	Information not available/ unknown
Education attainment of father	Edu_father	Numeric	1	No Schooling
			2	Kindergarten
			3	Primary (P1-P6)
			4	Lower Secondary (S1-S3)
			5	Upper Secondary (S4-S5 /S7/ Project Yi Jin/Yi Jin Diploma and craft level.)
			6	Post-secondary (Non-degree course)
			7	Post-secondary (Degree course or above)
			8	Others
			888	Information not available/ unknown
Education attainment of mother	Edu_mother	Numeric	1	No Schooling
			2	Kindergarten
			3	Primary (P1-P6)
			4	Lower Secondary (S1-S3)
			5	Upper Secondary (S4-S5 /S7/ Project Yi Jin/Yi Jin Diploma and craft level.)
			6	Post-secondary (Non-degree course)
			7	Post-secondary (Degree course or above)
			8	Others
			888	Information not available/ unknown
Housing type	Housing_type	Numeric	1	Public rental housing
			2	Subsidised housing
			3	Private housing

Variables	Proposed data structure			
	Variable name	Variable type	Values	Value labels
			4	Others
			888	Information not available/ unknown
Language at home	Language_at_ home	String	CHI	Chinese
			DUT	Dutch
			ENG	English
			PRT	Portuguese
			SNH	Sinhalese
			URD	Urdu
			ZAS	Other Asian and Oceanian Languages
			ZEU	Other European Language
			HND	Hindi
			RUS	Russian
			BNG	Bengali
			FRE	French
			GER	German
			IDN	Indonesian
			IRI	Irish
			ITA	Italian
			JPN	Japanese
			KOR	Korean
			NPL	Nepali
			PHL	Filipino
			SPA	Spanish
			VNM	Vietnamese
			THA	Thai
			PUN	Punjabi
			INA	Information not available/ unknown

Demographic Data Recorded by One B/Ds' System Only

21 categories of other subject-specific data involving 68 variables are found to be recorded by one B/Ds' system only. These data could be readily available and be provided directly to the consolidated children database without the need for data alignment.

Data Alignment

Six other child-related data that are recorded by multiple B/Ds are dynamic data. The key tasks involved in the data alignment are specified as follows:

Step 1 – B/Ds extract the six other child-related data (i.e. height, weight, education attainment of father, education attainment of mother, housing type and language at home) that are recorded by multiple B/Ds' systems, and other variables that are recorded by one B/Ds' system only.

Step 2 – B/Ds compile the extracted other child-related data for de-identification.

Step 3 – (Option 1) B/Ds transform the other child-related data (i.e. height, weight, education attainment of father, education attainment of mother, housing type and language at home) according to the proposed data structure and submit the de-identified data to the CDC Project Office.

(Option 2) B/Ds submit the de-identified data to the CDC Project Office, and the CDC Project Office transform the other child-related data (i.e. height, weight, education attainment of father, education attainment of mother, housing type and language at home) according to the proposed data structure.

While there could be two options on which party to transform the data, Option 1 is preferred as B/Ds have more understanding on the data structure of their systems and some re-coding may involve sensitive data for rare child cases. Taking also into consideration the issue of data sensitivity and confidentiality, B/Ds will be in a better position to conduct the transformation process of the data alignment.

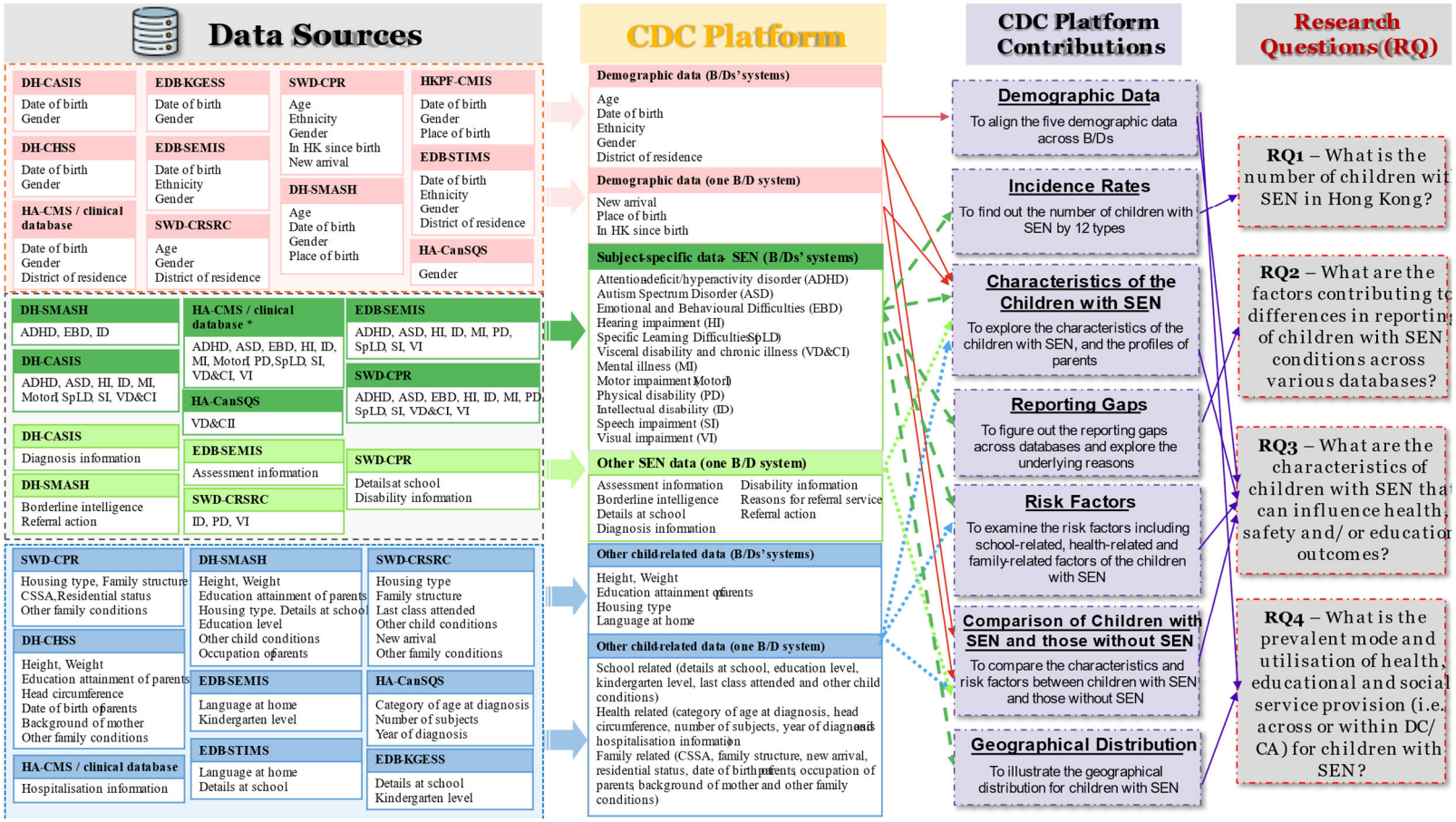
Step 4 – The CDC Project Office checks for the consistency of the data, and these data will be stored in the consolidated children database.

Step 5 – If trend monitoring is required, B/Ds will provide the data including the new entries of the children and updated dynamic data of the children who have already been in the consolidated children database to the CDC Project Office on a regular basis.

2.1.2.2.7 Anticipated Contribution of CDC Platform Upon Data Alignment

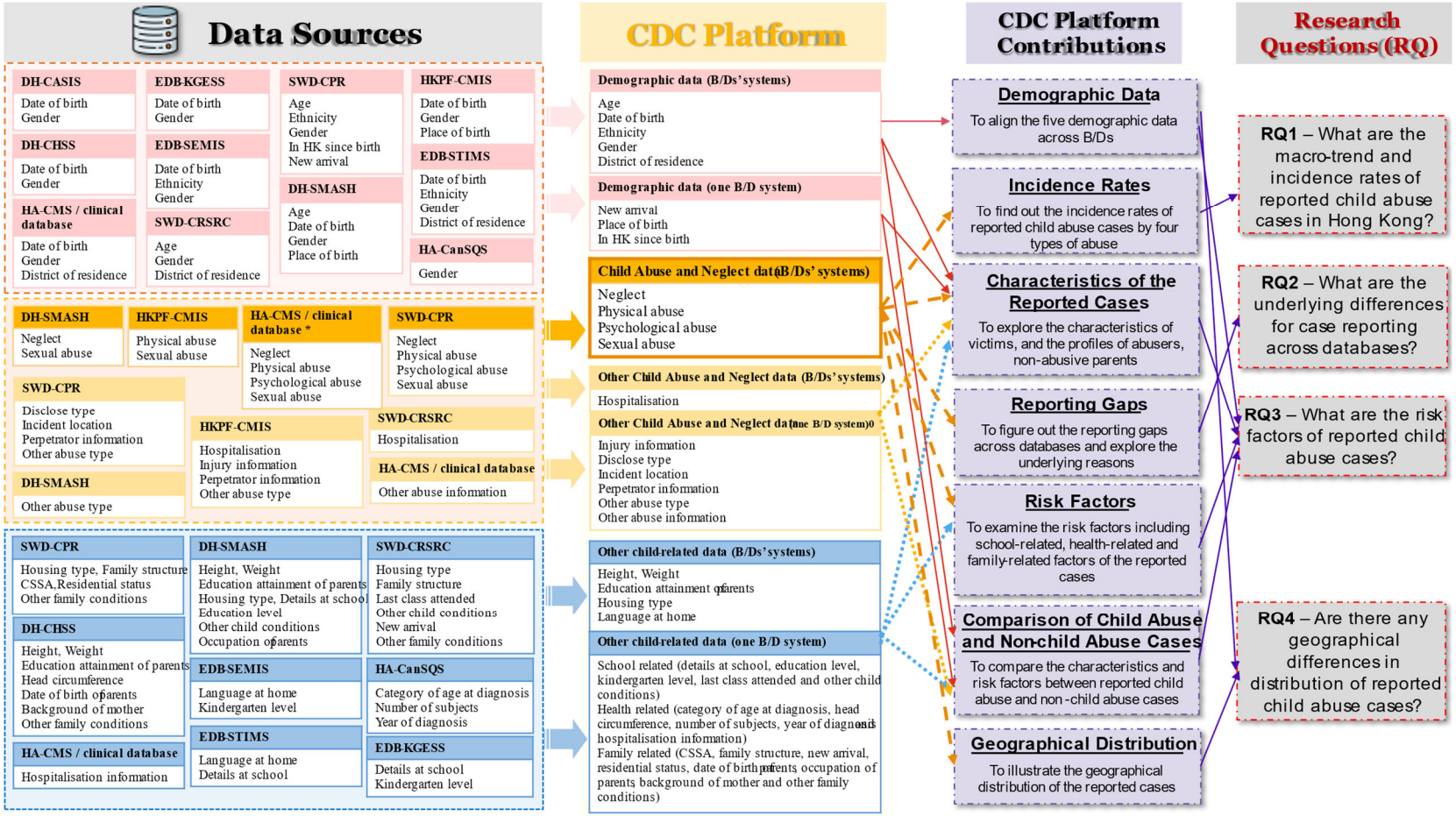
Figure 5 illustrates (i) the anticipated relations between data sources and the CDC platform; (ii) the corresponding contributions of the CDC platform on relevant demographic data, subject specific data and other child-related data; and (iii) how these contributions could inform some of the illustrative research questions set out in **Table 2** of **Section 2.1.1.1.3**, based on the above data alignment activities. An enlarged version of the diagram is available in **Appendix E**.

Figure 5 Anticipated Relations and Contributions of CDC Platform Upon Data Alignment



(*) the diagnoses are based on ICD-9-CM,with mapping to 2010 v ICD-10. Review on the diagnosis list by subject experts, which could take up to a few months, is recommended for a more comprehensive diagnosis list for child abuse & neglect, and SEN. The number of variables in the data category are therefore to be determined after the expert review.

Figure 5 (Cont'd) Anticipated Relations and Contributions of CDC Platform Upon Data Alignment



(*) the diagnoses are based on ICD-9-CM,with mapping to 2010 v ICD-10. Review on the diagnosis list by subject experts, which could take up to a few months, is recommended for a more comprehensive diagnosis list for child abuse & neglect, and SEN. The number of variables in the data category are therefore to be determined after the expert review.

2.1.2.2.8 Limitations, Challenges and Magnitude of Issue

Data Inconsistency

Data alignment is the process to transform the variables into the same data structure as far as possible for the purposes of conducting matching, comparison or cross-checking. Across systems, there may be instances of data inconsistency. For example, two systems may record a child as a girl whereas another system record the same child as a boy. In case of inaccurate or incomplete data, checking or verification of data could be done to rectify the data.

Liaison between B/Ds and the CDC Project Office may be required in order to handle cases of data inconsistency. However, if checking is not an option for some reasons, the consistent records of the majority systems may be adopted. The consolidated record will also be provided to the B/D for follow-up.

Data Structure to be Adjusted

The proposed data structure serves as the first step of the data alignment exercise by reading through the codebooks from B/Ds on these two priority areas. For the pilot projects or other research, the CDC Project Office and the appointed researchers should carefully study the codebooks and examine the data before the preparation of the data structure of the consolidated children database.

Besides, liaison between B/Ds and the CDC Project Office may be required for finalising the data structure and procedures required for the data alignment exercise.

Varied Data Collection Purposes and Definitions

B/Ds collect subject-specific data according to their own definitions or with reference to international codes. For instance, the International Classification of Disease, Ninth Revision, Clinical Modification (ICD-9-CM) currently adopted by HA is an official system of assigning codes to diagnoses and procedures associated with hospital utilisation.

Taking the variable “ADHD” as an example, one system records whether the child had high scores in “attention problem” and “hyperactivity” separately by adopting two different sets of instruments; another system records the child with ADHD after clinical assessment by doctor or psychologist. Although both systems consist of a variable measuring ADHD, the purposes of the data collection vary, hence, so does the methods of classifications.

Another example would be the severity level of intellectual disability. The severity level is recorded in different systems based on the assessment conclusion made by different specialists, including educational psychologists, clinical psychologists, psychiatrists, medical doctors, ophthalmologists, audiologists, speech therapists, etc. Since the assessments on the severity level of different B/Ds are based on various expertise, performing data alignment for this variable may not be feasible. In fact, it may be worth exploring the information on the severity level across systems to understand the assessments based on different specialists.

Under this circumstance, de-identified datasets from these different B/Ds should be made available to the researchers or parties commissioned to conduct the pilots, who would subsequently select the appropriate datasets to be adopted having regard to their research objectives. To illustrate this point, assuming the systems of EDB and DH both capture the severity level of intellectual disability in their system but two B/Ds rely on different specialists for the assessment, if this variable is requested by the researcher for the purpose of the pilot project, the CDC project office should present both datasets from DH and EDB to the researcher such that the researcher could decide on their own which datasets to adopt with regard to the research objectives. The CDC Office may present the respective datasets to the researcher under "two rows" (with relevant dataset details) for the ease of comparison.

As regards child physical abuse, two of the systems examined record cases related to physical abuse but another one records the health care after the reported physical abuse. Researchers may consider deriving the incidence rate of child physical abuse by consolidating the information or identifying gaps in case reporting across systems.

In sum, careful considerations should be given and review should be conducted prior to the use of subject-specific data across B/Ds' systems. Even for variables measuring the same construct and with aligned values, researchers should still examine the details of the variables provided by various B/Ds' systems before extracting valuable information.

Having set out the limitations and challenges in relation to the process of data alignment with regard to the systems examined, the number of records of the 11 systems of various B/Ds are tabulated below to indicate the potential volume of data which could be implicated by the data alignment exercise.

Table 17 Number of Records in B/Ds System

B/D	System	Number of Records[11]
SWD	CPR	20,168
	CRSRC	16,465
EDB	KGESS	332,175
	SEMIS	394,125
	STMIS	624,000
HA	CanSQS	1,927

[11] The numbers of records are based on information provided by B/Ds as of 30 November 2022. Where accumulative numbers of records are unavailable, average number of records per year are adopted instead.

B/D	System	Number of Records [11]
	CMS / clinical database	860,000
DH	CHSS	50,000
	SMASH	500,000
	CASIS	18,000
HKPF	CMIS	1,232
Total number of records		2,818,092

2.1.2.3 Data Alignment Plan

2.1.2.3.1 Summary of Proposed Data Items to be Aligned

Based on the database design for the two priority areas, a list of potential data items to be aligned are identified and summarised in **Table 18** below.

Table 18 List of Potential Data Items to be Aligned

Data item to be aligned	B/Ds involved	Data alignment	Remarks (e.g. mitigation measures, adjustments)
(1) Demographic data			
(1a) Demographic data recorded in multiple B/Ds' systems			
Age	DH, SWD	Data is aligned and recorded in numeric type (0-99)	The children age should be adjusted according to the system filing date of system.
Date of birth	DH, EDB, HA, HKPF, SWD	Data is transformed to string type in date format – DDMMYYYY	This standardised format should be adopted to ensure that the date of birth of the children across B/Ds' systems are aligned. This is useful for calculating the age of the children in case some B/Ds' systems do not capture the age as at the filing date.
Ethnicity	EDB, SWD	Data is transformed to numeric type (1-14)	To combine the detailed codes into the 14 proposed codes.
Gender	DH, EDB, HA, HKPF, SWD	Data is transformed to numeric type (1-3)	To change the variable type to numeric.
District of Residence	EDB, SWD	Data is transformed to numeric type (1-20)	To combine the detailed codes (say Yuen Long and Tin Shui Wai into Yuen Long district) into the 20 proposed codes.
(2) Subject-specific data – Child Abuse			
(2a) Key subject-specific data – Child Abuse			
Neglect	DH, HA(*), SWD	Data is transformed to numeric type (0,1)	• For DH system – to change the value labels to 1=yes and 0=no. • For SWD system – to recode the data for 1=yes if Abuse Type = 2 (neglect); and for 0=no if neglect is not selected.
Physical abuse	HA(*), HKPF, SWD	Data is transformed to numeric type (0,1)	• For HKPF system – to recode the data for 1=yes if Child Abuse (physical crimes); and 0=no if otherwise. • For SWD system – to recode the data for 1=yes if Abuse Type = 1 (physical abuse); and for 0=no if physical abuse is not selected.
Psychological abuse	HA(*), SWD	Data is transformed to numeric type (0,1)	• For SWD system – to recode the data for 1=yes if Abuse Type = 4 (psychological abuse); and for

Data item to be aligned	B/Ds involved	Data alignment	Remarks (e.g. mitigation measures, adjustments)
			0=no if this psychological abuse is not selected.
Sexual abuse	DH, HA(*), HKPF, SWD	Data is transformed to numeric type (0,1)	• For DH system – to change the value labels to 1=yes and 0=no. • For HKPF system – to recode the data for 1=yes if Child Abuse (sexual crimes); and 0=no if otherwise. • For SWD system – to recode the data for 1=yes if Abuse Type = 3 (sexual abuse); and for 0=no if sexual abuse is not selected.
(2b) Other subject-specific data recorded in multiple B/Ds' system – Child Abuse			
Hospitalisation	HKPF, SWD	Data is transformed to numeric type (0,1)	• For HKPF system to change the variable type to numeric. • For SWD system – to recode the data for 1=yes if DW=1 or WC = 1 or WR = 1; and for 0=no if otherwise.
(2) Subject-specific data – SEN			
(2a) Key subject-specific data – SEN			
Attention-deficit/ hyperactivity disorder (ADHD)	DH, EDB, HA(*), SWD	Data is transformed to numeric type (0,1)	• For DH systems – to change the variable type to numeric. • For EDB system – to recode the data for 1=yes if SEN_Type 1-5 = ADHD; and for 0=no if otherwise. • For SWD system – to change the variable type to numeric.
Autism Spectrum Disorder	DH, EDB, HA(*), SWD	Data is transformed to numeric type (0,1)	• For DH systems – to change the variable type to numeric. • For EDB system – to recode the data for 1=yes if SEN_Type 1-5 = PERDE; and for 0=no if otherwise. • For SWD system – to change the variable type to numeric.
Emotional and Behavioural Difficulties	DH, HA(*), SWD	Data is transformed to numeric type (0,1)	• For DH systems – change the variable type to numeric. • For SWD system – to change the variable type to numeric.
Hearing impairment	DH, EDB, HA(*), SWD	Data is transformed to numeric type (0,1)	• For DH systems – change the variable type to numeric. • For EDB system – to recode the data for 1=yes if SEN_Type 1-5 = HI; and for 0=no if otherwise.

Data item to be aligned	B/Ds involved	Data alignment	Remarks (e.g. mitigation measures, adjustments)
			• For SWD system – to change the variable type to numeric.
Intellectual disability	DH, EDB, HA(*), SWD	Data is transformed to numeric type (0,1)	• For DH systems – change the variable type to numeric. • For EDB system – to recode the data for 1=yes if SEN_Type 1-5 = ID; and for 0=no if otherwise. • For SWD systems – to change the variable type to numeric.
Mental illness	DH, EDB, HA(*), SWD	Data is transformed to numeric type (0,1)	• For DH system – to recode the data for 1=yes if Developmental problems – Anxiety problem / disorders =15-01 to 15-10; and for 0=no if otherwise. • For EDB system – to recode the data for 1=yes if SEN_Type 1-5 = MI or PSYPR; and for 0=no if otherwise. • For SWD system – to change the variable type to numeric.
Motor impairment	DH, HA(*)	Data is transformed to numeric type (0,1)	• For DH systems – change the variable type to numeric.
Physical disability	EDB, HA(*), SWD	Data is transformed to numeric type (0,1)	• For EDB system – to recode the data for 1=yes if SEN_Type 1-5 = PD; and for 0=no if otherwise. • For SWD systems – to change the variable type to numeric.
Specific Learning Difficulties (SpLD)	DH, EDB, HA(*), SWD	Data is transformed to numeric type (0,1)	• For DH systems – change the variable type to numeric. • For EDB system – to recode the data for 1=yes if SEN_Type 1-5 = SPLD; and for 0=no if otherwise. • For SWD system – to change the variable type to numeric
Speech impairment	DH, EDB, HA(*), SWD	Data is transformed to numeric type (0,1)	• For DH systems – change the variable type to numeric. • For EDB system – to recode the data for 1=yes if SEN_Type 1-5 = SLI; and for 0=no if otherwise. • For SWD system – to change the variable type to numeric.
Visceral disability and chronic illness	DH, HA(*), SWD	Data is transformed to numeric type (0,1)	• For DH system – change the variable type to numeric.

Data item to be aligned	B/Ds involved	Data alignment	Remarks (e.g. mitigation measures, adjustments)
			• For HA – CanSQS – to recode the data for 1=yes if site group = 11 to 21; and for 0=no if otherwise. • For SWD system – to change the variable type to numeric.
Visual impairment	DH, EDB, HA(*), SWD	Data is transformed to numeric type (0,1)	• For DH systems – change the variable type to numeric. • For EDB system – to recode the data for 1=yes if SEN_Type 1-5 = VI; and for 0=no if otherwise. • For SWD systems – to change the variable type to numeric.
(3) Other child-related data			
(3a) Other child-related data recorded in multiple B/Ds' systems			
Height	DH	Data is recorded in numeric type (in cm)	No data alignment is required.
Weight	DH	Data is recorded in numeric type (in kg)	No data alignment is required.
Education attainment of Father	DH	Data is transformed to numeric type (1-8)	For DH systems – to change the variable type to numeric.
Education attainment of Mother	DH	Data is transformed to numeric type (1-8)	For DH systems – change the variable type to numeric.
Housing Type	DH, SWD	Data is transformed to numeric type (1-4)	For DH system and SWD systems – to combine the detailed codes (say Home Ownership Scheme and Tenants Purchase Scheme into 56 anonymized housing) into the 4 proposed codes.
Language at Home	EDB	Data is recorded in string type	No data alignment is required.

(*) For HA CMS / clinical database – the diagnoses are based on ICD-9-CM, with mapping to 2010 v ICD-10. Review on the diagnosis list by subject experts, which could take up to a few months, is recommended for a more comprehensive diagnosis list for child abuse & neglect, and SEN. The number of variables in the data category are therefore to be determined after the expert review.

2.1.2.3.2 General Steps for Alignment

As demonstrated in **Section 2.1.2.2** above, children data are collected and recorded using different data structures[12] in the systems of various B/Ds. A piece of information measuring the same aspect of a child could be recorded using different variable labels, variable types, values and value labels in the systems of various B/Ds to suit their specific user and operating requirements when these systems were designed and developed. For the purpose of compiling a consolidated children database, data alignment should be conducted in this light to gather the data from B/Ds' systems.

To this end, three stages of data alignment detailing the general steps for alignment are proposed in **Table 19** below. A diagram illustrating the steps involved in these three stages is set out at **Appendix E**.

Table 19 Three Stages of Data Alignment

Stages	Steps	Actions	Parties Involved
Stage 1 – Preparation Work	Step 1	The relevant WG under CoC overseeing CDC decides a priority area (or a specific topic), then the Data Advisory Unit of the CDC Project Office proposes a research framework. During this step, decision should be made on whether the research project is a one-off project or a trend-monitoring one to be conducted on a regular basis.	WG, CDC Project Office
	Step 2	The relevant B/Ds examine the data of the systems they managed that are related to the proposed research framework in consultation with the Data Advisory Unit of the CDC Project Office; and discussions or meetings may be required with the relevant professionals in the B/Ds.	B/Ds, CDC Project Office
	Step 3	The relevant B/Ds provide a codebook for the list of variables according to the results of Step 2 to the CDC.	B/Ds
	Step 4	The Data Analytics and Linkage Unit of the CDC Project Office examines the codebooks, check the data structures, and classify the variables into three main categories.	CDC Project Office
	Step 5	The Data Analytics and Linkage Unit prepares a codebook of a consolidated children database and sets out the proposed data structures for the variables that are recorded by multiple B/Ds' systems after discussions with B/Ds.	CDC Project Office
Stage 2 – Data Provision and Checking	Step 1	B/Ds extract the relevant data.	B/Ds
	Step 2	B/Ds conduct the de-identification of the extracted data.	B/Ds
	Step 3	Option 1	B/Ds,

[12] Data structures: "Variable name" refers to the name assigned to each variable; "Variable label" refers to a brief description of each variable; "Values" refers to the actual coded values in the data for each variable; "Value labels" refers to the textual descriptions of the coded values; "Variable types" refers to the types of each variable, i.e. numeric, string; and "Definitions" refers to definitions or international classifications of the variable if any.

Stages	Steps	Actions	Parties Involved
		B/Ds transform the data according to the proposed data structure and submit the de-identified data to the Data Analytics and Linkage Unit of the CDC Project Office. Option 2 B/Ds submit the de-identified data to the CDC Project Office, and the CDC Project Office transform the data according to the proposed data structure. While there could be two options on which party to transform the data, Option 1 is preferred as B/Ds have more understanding on the data structure of their systems and some re-coding may involve sensitive data for rare child cases. Taking also into consideration the issue of data sensitivity and confidentiality, B/Ds will be in a better position to conduct the transformation process of the data alignment.	CDC Project Office
	Step 4	The Data Analytics and Linkage Unit checks for consistency of the demographic data provided by the B/Ds. For demographic static data, if the data are matched, new variables will be created and stored in the consolidated children database; and if the data are not matched, the Data Analytics and Linkage Unit will inform B/Ds to counter check the details. For demographic dynamic data, the latest records from the B/Ds will be retrieved and stored in the new variables of the consolidated children database.	CDC Project Office
	Step 5	B/Ds counter check the details of those unmatched demographic static data and provide the updated data to the Data Analytics and Linkage unit.	B/Ds
	Step 6	The Operations Unit of the CDC Project Office updates the data in the consolidated children database. Step 4 to Step 6 should be repeated until the demographic static data are matched or in some circumstances such as unavailable information and incomplete checking, consensuses have been made across B/Ds on the inconsistent demographic static data to be recorded in the consolidated children database. Discussions or meetings between B/Ds and the CDC Project Office may be required for handling the data inconsistences.	CDC Project Office, B/Ds
	Step 7	The Data Analytics and Linkage Unit checks for the consistency of the subject-specific and other child-related data provided by the B/Ds, and these data will be stored in consolidated children database. Where appropriate, the CDC Project Office derives new variables for the subject-specific data according to various research hypotheses.	CDC Project Office
Stage 3 – Data Updating	Step 1	If trend monitoring is required, B/Ds will provide the data including the new entries of the children and	B/Ds

Stages	Steps	Actions	Parties Involved
		updated dynamic data of the children who have already been in the consolidated children database to the CDC Project Office on a regular basis.	
	Step 2	The Data Analytics and Linkage Unit and B/Ds review the data structures of the variables from time to time. If required, B/Ds consider doing the adjustments on data collection or as when system will be upgraded in the future.	CDC Project Office, B/Ds
	Step 3	The Operations Unit maintains the consolidated children database.	CDC Project Office

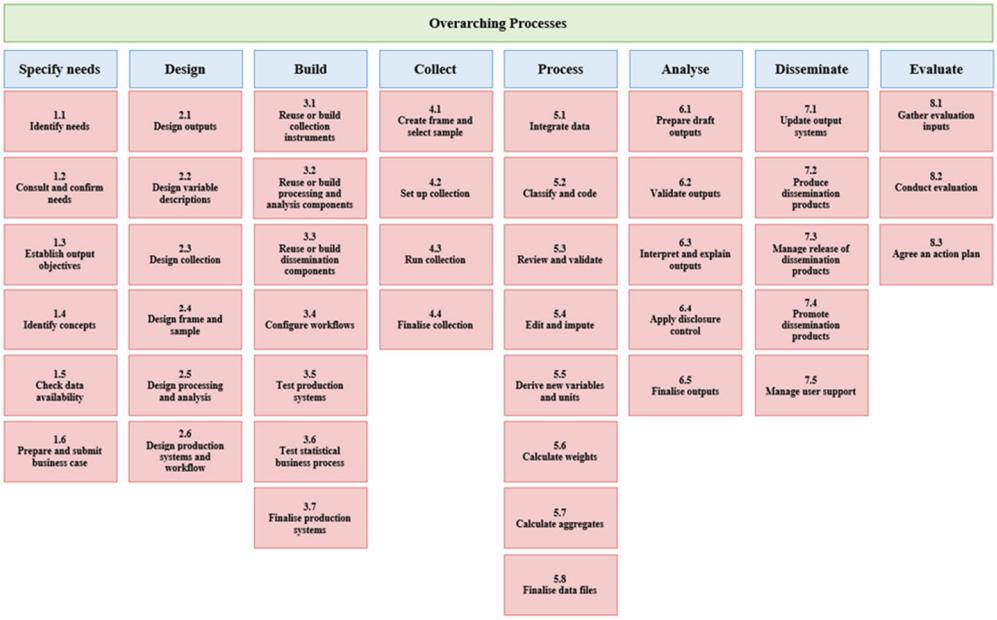
2.1.2.4 Data Management and Governance

2.1.2.4.1 Data Lifecycle for Statistical Data, Process and Instruments of CDC

It is important to understand and plan for the data lifecycle within CDC in order to manage the data being collected, processed and disseminated, as well as associated information, under the proposed Data Governance and Management Framework of CDC. The UNECE/Eurostat/OECD Generic Statistical Business Process Model (GSBPM)[13] could be referenced for adoption to meet the specific needs of CDC operations and associated data lifecycle. GSBPM is an international standard to describe statistical processes in a statistical organisation using a common language, and has been adopted by National Statistical Offices around the world. The overarching data lifecycle activities/processes described in GSBPM include (i) Specify Needs; (ii) Design; (iii) Build; (iv) Collect; (v) Process; (vi) Analyse; (vii) Disseminate and (viii) Evaluate.

Together, these processes constitute the essential activities from conceptualisation to dissemination of data throughout the management of statistical data and research outcomes of social science studies within CDC. The below figure is a linear presentation of the GSBPM processes. It is presented as a matrix but there are many possible means to traverse the GSBPM processes based on the specific data lifecycle requirements of the organisations, including iterative loops within and between processes etc.

Figure 6 GSBPM Processes



[13] https://statswiki.unece.org/display/GSBPM/GSBPM+v5.1

The overarching data lifecycle activities/processes described in GSBPM are explained as follows:

1. Specify Needs

This is the beginning of the data lifecycle where a need for new statistics is identified, or feedback about current statistics initiates a review. It includes various activities associated with identifying detailed statistical needs, proposing high level solution options and preparing business cases to meet these needs. In CDC's context, this could be initiated by a specific policy initiative, a thematic research topic, or a project proposed by the academic.

2. Design

This includes all the design elements needed to define or refine the statistical products (i.e. data) required for a business case. This key process specifies all relevant metadata, ready for use later in the statistical business process, as well as quality assurance procedures. The activities involved in the Design process can be conducted outside of CDC.

3. Build

During this process, different statistical tools and methodologies, such as survey forms, statistical returns, etc., are built and tested based for use for data collection in the 'live' environment. The outputs of the Design process influence the selection of reusable processes, instruments, and information that are assembled and configured to create the complete environment for subsequent processes. The activities involved in the Build process can be performed outside of CDC.

4. Collect

This process involves the collection of necessary information (data and metadata) through different collection modes (including extractions from statistical, administrative and other non-statistical registers and databases), and loads them into the appropriate environment for further processing. This process will be carried out through the provision of tools and functions available in CDC.

5. Process

The main objective of this process is primarily the cleansing and preparation of collected data for analysis including activities such as checking, cleaning, and transforming data, such that they can be analysed and disseminated as statistical outputs. This process will be carried out through the provision of data processing mechanism in CDC.

6. Analyse

This process involves the production of statistical outputs by examining in detail and making ready for dissemination through the preparation of statistical content (including commentary, technical notes, etc.), and ensuring outputs are 'fit for purpose' prior to dissemination. This process will be carried out through the provision of data analytic tools available in CDC.

7. Disseminate

This process manages the release of the statistical products through different activities associated with assembling and releasing a range of static and dynamic products

(aggregated or consolidated data) via a range of channels with the objective to support other parties to access and use the statistical outputs being released. This process will be carried out through the data dissemination mechanism available in CDC.

8. Evaluate

Although this process takes place at the end of the instance of the process, it relies on inputs gathered throughout the entire data lifecycle by drawing on a range of quantitative and qualitative inputs, and identifying and prioritising potential improvements. The activities involved in this process is typically conducted outside of CDC.

Furthermore, the data lifecycle of CDC should also involve the management of the surveys and statistical returns being prepared, i.e. the instrument to facilitate the collection of data. The Data Documentation Initiative ("DDI") is another international standard for lifecycle management of surveys and questionnaire information, statistical returns and research data for social science studies. DDI has been adopted by National Statistical Offices including Australia, Canada and New Zealand, and is currently widely promulgated by the European Commission among its national institutes of statistics. **Figure 7** is a high-level diagram of the DDI. To ensure complete coverage of the data and metadata being processed in CDC, DDI should also be considered for inclusion as a standard for adoption in managing data lifecycle within CDC.

Figure 7 High-level Diagram of the DDI[14]

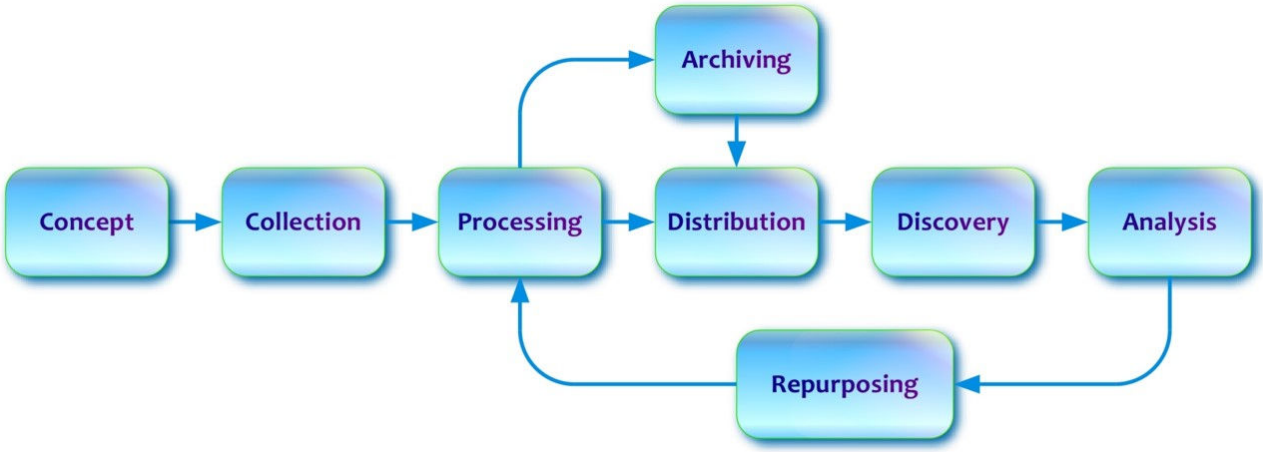


Table 20 presents the mapping between the GSBPM and DDI from a data lifecycle perspective applicable to CDC:

[14] Source: https://ddi-lifecycle-documentation.readthedocs.io/en/latest/User%20Guide/Introduction.html

Table 20 Mapping between the GSBPM and DDI

GSBPM (Statistical data and processes)	DDI (Survey, questionnaire and Statistical returns)
1. Specify Needs	• Concept • Repurposing
2. Design	
3. Build	
4. Collect	• Collection
5. Process	• Processing • Repurposing
6. Analyse	• Discovery • Analysis • Processing
7. Disseminate	• Distribution • Archiving
8. Evaluate	N/A

In addition to GSBPM and DDI, for data quality, although the IMF Data Quality Assessment Framework (DQAF)[15] is not designed specifically for social statistics, the generic framework of DQAF could be considered and tailored for CDC since it addresses the overall data quality aspect through coverage in institutional environments, statistical processes, and characteristics of the statistical products.

Data Retention and Disposal

For the handling of data retention and disposal in CDC, considerations need to be given to the (i) data Sources; (ii) data processed by CDC; and (iii) data Disseminated by CDC.

For data collected by CDC from various data sources, such as those originating from B/Ds or NGOs, the retention and disposal of the source data residing in B/Ds or NGOs should follow the corresponding policy and procedures of the corresponding organisations. For B/Ds, the overarching disposal schedule by Government Records Service (GRS) of disposing administrative and programme records after seven-year of retention should be observed. B/Ds can further tailor the data retention periods of certain records beyond the seven-year duration based on their specific administrative

[15] https://dsbb.imf.org/dqrs/DQAF

and operational needs with appropriate justifications. Similar practices are expected to be observed by NGOs as well. For disposal of personal data, B/Ds and NGOs should observe and comply with Section 26 of the Personal Data (Privacy) Ordinance ("PDPO") and Data Protection Principle ("DPP")2(2) in Schedule 1 to the PDPO, B/Ds and NGOs should have a personal data erasure policy in place on how different types of personal data records, regardless physical or digital, are being handled. Data being contributed to and collected by CDC should be bound by the personal data erasure policy of individual organisations.

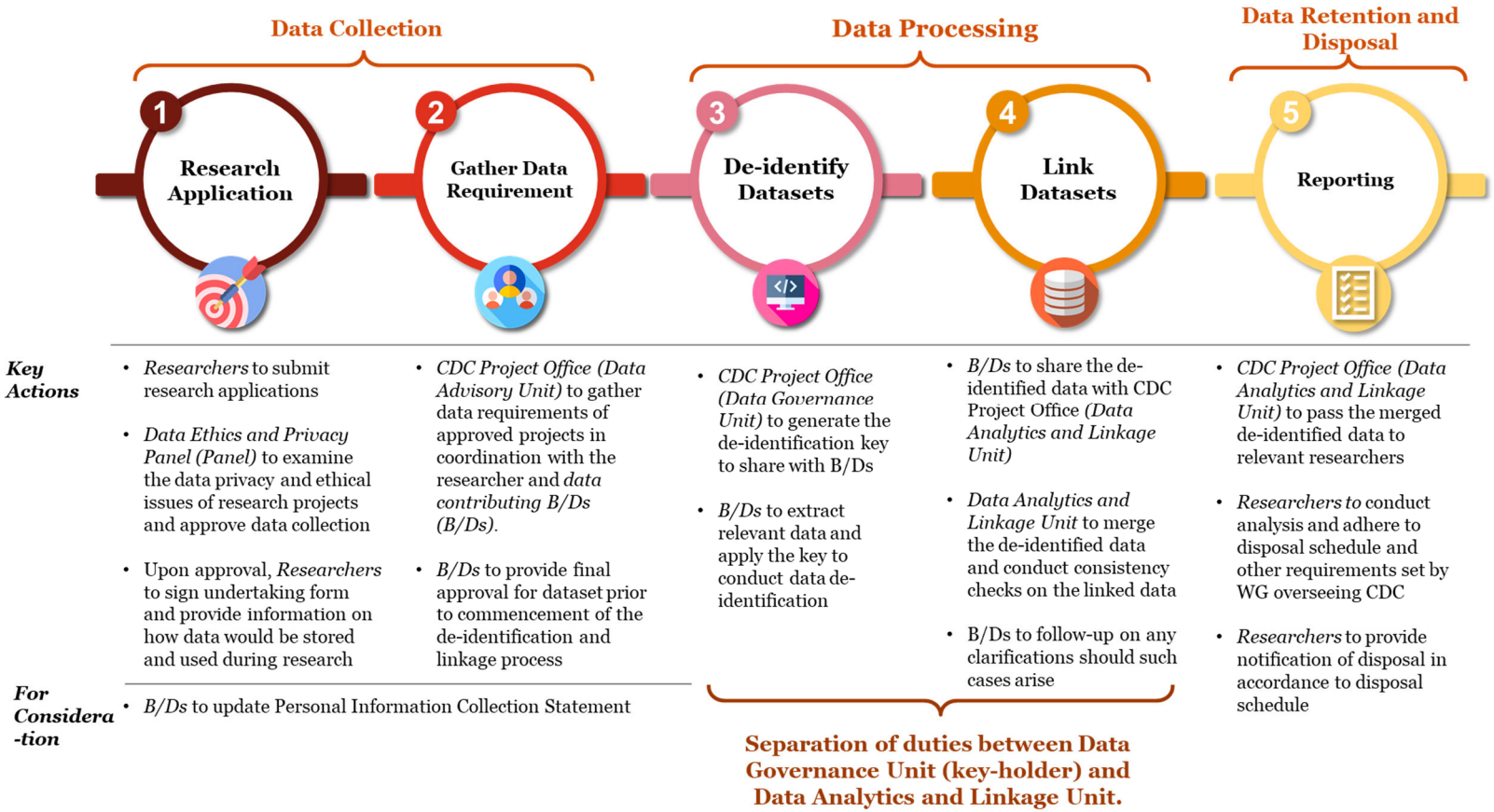
For data being processed by CDC, in particular personal data related information being collected from various data sources, the personal data collected should only be retained no longer than necessary for the fulfilment of the purpose of which such data is or is not being used for processing and analysis. The retention period of the collected data in CDC for processing and analysis must be defined for each individual projects, and must be approved by the proposed Data Ethics and Privacy Panel of CDC before the commencement of the individual projects. The disposal of the source data should be pragmatically followed and the Data Ethics and Privacy Panel would have the right to audit selected projects if necessary.

For data disseminated by CDC, to cater for longitudinal studies of children from infancy to adulthood, it is necessary to maintain the disseminated data for at least 18 years (covering from birth to becoming adult). Considering the disseminated data would be classified as records under the Hong Kong Government general practice, the disposal schedule of GRS should also be observed. Thus, an overall retention period of 25 years (18 years plus 7 years) would be applied to all disseminated data from CDC before disposal or further archiving.

2.1.2.4.2 Key Parties and Procedures for Data Governance Concerning Data De-identification and Linkages

The key parties involved and procedures of approving and conducting data de-identification and data linkages, by various phases in the data life cycle, are illustrated below.

Figure 8 Approval Process for Data De-identification and Data Processing



Data collection

For the purpose of conducting pilot projects under the Foundational Mode or data linkage projects under the Enhancement Mode, the Data Ethics and Privacy Panel must endorse and approve the collection of data. Potential criteria for assessing research applications may include whether the use of data serves the public good, whether confidentiality and data security are ensured, whether the research methods are consistent with recognised standards of integrity and quality, as well as the legal compliance, public views, and transparency of the data used and method employed in the research. Taking reference from the experience of the National Statistician's Data Ethics Advisory Committee (NSDEAC) of UK[16], an ethics self-assessment tool could be developed based on the agreed criteria to be adopted by the Panel, to serve as an easy-to-use framework to help researchers review the ethics implications of their projects throughout the research cycle prior to the formal review by the Panel.

Upon approval of the application, researchers would need to sign undertaking forms concerning the use of the requested data. This would be important especially if the researcher is allowed to make multiple data requests for separate projects. There is a higher risk of identifying individual data subjects if a researcher is able to gain access to multiple data sets of the same group of data subjects. By signing the undertaking forms, the researcher declares that he or she (i) would not use the data requested for different projects and (ii) would provide notification for the disposal of data. Alongside the declaration, the undertaking forms would also request from the researcher detailed information on how data would be stored and used during their research.

It is to note that for data linkage projects, the collection of existing de-identified data from relevant B/Ds may not require retrospective consent from data subjects as this does not constitute a non-compliance of the Personal Data (Privacy) Ordinance.[17] It is recommended that both the Department of Justice (DOJ) and the PCPD should be consulted for a more detailed assessment on the privacy and legal implications for such retrospective sharing of de-identified data. Moving forward, to instil public trust, it is recommended that the Personal Information Collection Statement (PICS) adopted by relevant B/Ds be updated in anticipation of future data sharing opportunities with CDC. Further elaboration on PICS is at **Section 2.1.2.4.4.**

Data processing

For conducting data linkage projects, data required from relevant B/Ds will be identified upon confirmation of the research topic with support from the CDC Project Office. There will be a segregation of functions and a Chinese wall whereby two

[16] UK Statistics Authority, National Statistician's Data Ethics Advisory Committee - Ethics Self-Assessment Tool <https://uksa.statisticsauthority.gov.uk/the-authority-board/committees/national-statisticians-advisory-committees-and-panels/national-statisticians-data-ethics-advisory-committee/ethics-self-assessment-tool/>

[17] Based on consultation with Mr Stephen Lau, JP, Senior Adviser at PwC Hong Kong and the former Privacy Commissioner for Personal Data for Hong Kong.

separate divisions within the CDC Project Office will be responsible for generating the de-identification key and for conducting the data linkage respectively. The Data Governance Unit within the IT and Data Governance Division will be responsible for generating the de-identification key, which will be shared with data contributing B/Ds for conducting the actual de-identification process. Following which, the de-identified data will be shared with the Data Analytics and Linkage Unit within the Data Advisory and Analytics Division of the CDC Project Office for conducting the linkage. The merged de-identified data will then be passed on to relevant researcher for conducting further analysis. Further details of the mechanism for conducting data de-identification and linkage are as follows:

- **Project-specific de-identification key/conversion algorithm:** A project-specific set of conversion algorithm will be used to de-identify data and enable linkages[18] to safeguard personal data of children required for research. For cross-sectional linkage projects, different set of Pseudo Identifiers could be applied to different projects even for the same set of administrative data. This would prevent the possibility of linking data requested for different projects, thereby reducing the risks of identifying individual data subject. For longitudinal data linkage projects, different parameter of the de-identification key can be applied whenever new sets of administrative data would need to be included in the longitudinal study. It would be necessary to apply a different key for the de-identification process whenever new sets of administrative data would be included in the longitudinal study. This one-off key should be applied to all previous administrative data in the past under the same longitudinal study rather than just the new set of administrative data alone. With this approach, the administrative data over the specific period is de-identified using the same key within each year. However, this set of de-identified data cannot be cross-referenced with the set of de-identified data from previous years as different keys are being used.
- **De-identification key to be outside of the CDC platform:** The de-identification key/conversion algorithm will be operated outside of the CDC platform and be held by the Data Governance Unit of the CDC Project Office. The CDC Project Office shall ensure segregation of duties between the function of generating the de-identification key (i.e., the Data Governance Unit) and the function for conducting data linkage of de-identified data (i.e., the Data Analytics and Linkage Unit), with only designated personnel of relevant B/Ds gaining access to the key. There shall also be physical separation of these two teams and respective IT infrastructure. Upon receipt of the de-identification key, the designated

[18] It is to note that the case of the South Australian Early Childhood Development Project, the conversion algorithm and linkage function are both provided by SA-NT. SA-NT would hold the master key and pass onto relevant data custodians to convert personal information to unique identifiers. However, in the local context, specifically with reference to a data matching exercise on people with disabilities and chronic diseases conducted in Hong Kong, the conversion algorithm or the master key was held by a separate party (e.g. Hospital Authority) from that undertaken data linkage (i.e. C&SD).

personnel of the B/Ds will convert the identifiable dataset to de-identified data with the de-identification key, where personal information, such as HKID number, will be replaced by a set of unique identifiers. The de-identified datasets will then be combined and matched based on the unique identifier by a separate section of the de-identification and data linkage body. This will then be provided to the requestor for further processing and analysis. The identification process shall also be subject to monitoring and oversight by the Data Ethics and Privacy Panel.

Data Retention and Disposal

It is envisaged that procedures shall be established for data retention and disposal. In particular, the disposal schedule (or the retention period) of data will be determined by the sensitivity of data being collected and for fulfilling the purpose of collecting such data, with consideration to the sources, processing, and dissemination of data. Therefore, a general disposal schedule is to be defined by the relevant CoC Working Group overseeing CDC, with exceptions to be considered on a project-by-project basis.

The researchers shall follow these procedures in accessing and using CDC's data. By retaining data for a specified period, research can be verified, if necessary, in accordance with undertakings given to data providers. The CDC Data Ethics and Privacy Panel either endorses the proposed retention period for data collections or may choose to stipulate alternative data retention requirements.

Datasets that are no longer used shall be indicated as "inactive". These include data that are kept in accordance with their retention requirements. Access privileges of all users will be reviewed by data custodians. Data should be destroyed if it has reached its minimum retention period and is no longer required for continuing business/ research or accountability purposes. In accordance with the undertaking form signed upon research approval, the CDC would be notified by the researcher on the data disposal once conducted.

2.1.2.4.3 Governance on Overall Data Alignment

Given that it will take time to identify appropriate set of "key data" and for B/Ds to adhere to any agreed practices/guidelines, an incremental approach shall be adopted for data alignment. Key tasks to be conducted and the responsible parties involved under the incremental approach are set out in **Table 21**. A key principle for this approach is to ensure that there is minimal disruption to the operations of B/Ds while ensuring that new data alignment requirements will be duly considered during new development and/or revamp of IT systems. In short, a set of data standards and guidelines shall be formulated as the guiding principles to facilitate the identification of key data for alignment in the long run.

At the start of CDC development, it is anticipated that the execution of the pilot projects on the two priority areas will be key in providing recommendations on data to be prioritised for alignment. These recommendations will then be considered by the relevant WG overseeing the CDC following the conclusion of the pilot projects. A preliminary set of criteria could be considered to facilitate identification of appropriate data type as "key data". For instance, "key data" shall generally exhibit the following characteristics:

(1) Generally required for understanding children's well-being (across data domains);

(2) Potential for enhanced efficiency for research and analysis;

(3) Factual and not subject to judgmental interpretation; and

(4) Result in minimal interruption to B/Ds' operations.

In general, it is proposed that an incremental approach should be used for data alignment. Giving due consideration to the proposed governance structure as detailed in **Section 2.1.1.3, Table 21** presents the proposed key objectives, key tasks and corresponding drivers for alignment in the short term and medium to long term.

Table 21 Incremental Approach for Data Alignment

Short Term – Key Objectives**:**
(1) To formalise a mechanism for identifying potential "key data" for alignment;
(2) To commence the process of setting data standards/guidelines for most prioritised "key data"; and
(3) To advise on data conversion function.

Key Tasks		Key Driver
1.	Assign a relevant WG under CoC to provide oversight to CDC, including advising on "key policy areas" to be prioritised for data alignment and analysis, and overseeing the setting of data standards and guidelines for alignment (See functions of the concerned WG at Section 2.1.1.3)	CoC
2.	Commission the pilot project for the first priority area of "Children with Risk of Abuse and Neglect", in which the engaged parties will be required to provide recommendations on data to be prioritised for alignment	CoC
3.	Identify "key policy areas" to be prioritised for data alignment and analysis taking into consideration the strategic priorities of CoC, and recommendations arising from the first pilot project	Relevant CoC WG
4.	Oversee the formulation of data standards/guideline including usage of data under the purview of research ethics with reference to local and overseas best practices	Relevant CoC WG
5.	Commission the pilot project for the second priority area of "Children with Special Educational Needs"	CoC
6.	Perform task 3 and 4 by applying the experience of the first pilot project	Relevant CoC WG
7.	Formulate data standards/guidelines for the first set of "key data" in consultation with B/Ds	CDC Project Office
8.	Provide data alignment/conversion support for pilot projects in the priority areas, and document any agreed data alignment practices among B/Ds over time	CDC Project Office

Medium to Long Term – Key Objectives:
(1) To monitor the progress of data alignment and;
(2) To provide continued support for data conversion and alignment

Key Tasks		Key Driver
1.	Monitor the progress of data alignment and overseeing the formulation of any new standards/guidelines such as project-specific practices/procedures for topics of concern that are investigated over time	CDC Project Office
2.	Continue to provide data alignment/conversion support for data linkage projects and provide reporting to the relevant WG overseeing CDC on potential practices that could be further aligned for specific topics of concern	CDC Project Office
3.	Identify new policy areas for data alignment and analysis, and advise B/Ds on considering new data alignment requirements with reference to data alignment exercises conducted during new development and/or revamp of B/Ds' IT systems where appropriate	Relevant CoC WG

2.1.2.4.4 Personal Information Collection Statement (PICS)

It is to note that the CDC's proposed use of anonymised data shall not constitute non-compliance of the PD(P)O.

For legal compliance with the notification requirements under the Data Protection Principle (DPP) 1(3) and DPP5 of the PD(P)O, the personal data collection process by the data user is preceded or accompanied by the issuance or display to the data subject of a PICS[19].

Under the prevalent framework, anonymisation of personal data is one of the general measures that allows the sharing of personal data for purposes excluded or different from that specified in the PICS for the original data collection. If the personal data held is anonymised to the extent that the data user or anyone else will not be able to directly or indirectly identify the individuals concerned, the data is not considered to be personal data under the Ordinance, as stated in the guidance document "Personal Data Erasure and Anonymisation".

Nonetheless, further consultation with DoJ and PCPD shall be required concerning the data linkage projects which involve the sharing of de-identified data.

To further instil trust in our community, additional terms for PICS in relation to data sharing with CDC shall be developed for common use by all departments in all their prevalent PICS during the data collection process for various purposes.

Based on our consultation with a subject matter expert on privacy of personal data, a draft of the additional statement has been supplemented below for illustrative purposes, which shall be subject to consultation with the Privacy Commissioner of Personal Data (PCPD) upon decision to adopt during implementation:

For illustrative purposes only

"Your personal data may be disclosed and shared with other relevant government departments for statutory requirements; to attain purpose that is benefits to you; and to serve the public interest.
In considering such data disclosure and sharing, we would perform a privacy impact assessment with regard to the inherent sensitivity of your data, the possible harm to you and your reasonable expectations in the use of your data. Based on the assessment, appropriate measures would be taken to mitigate such potential impact, e.g. anonymisation of your personal data for disclosure so that you are no identified (please note that suitably anonymised data would not be regarded legally as personal data), where appropriate restricted access of the data only by relevant departmental units on a "need to know/ access" basis, and monitoring of proper data use after data disclosure."

[19] Office of the Privacy Commissioner for Personal Data, Hong Kong, Guidance on Preparing Personal Information Collection Statement and Privacy Policy Statement
<https://www.pcpd.org.hk/english/publications/files/GN_picspps_e.pdf>

2.1.3. Technology Dimension

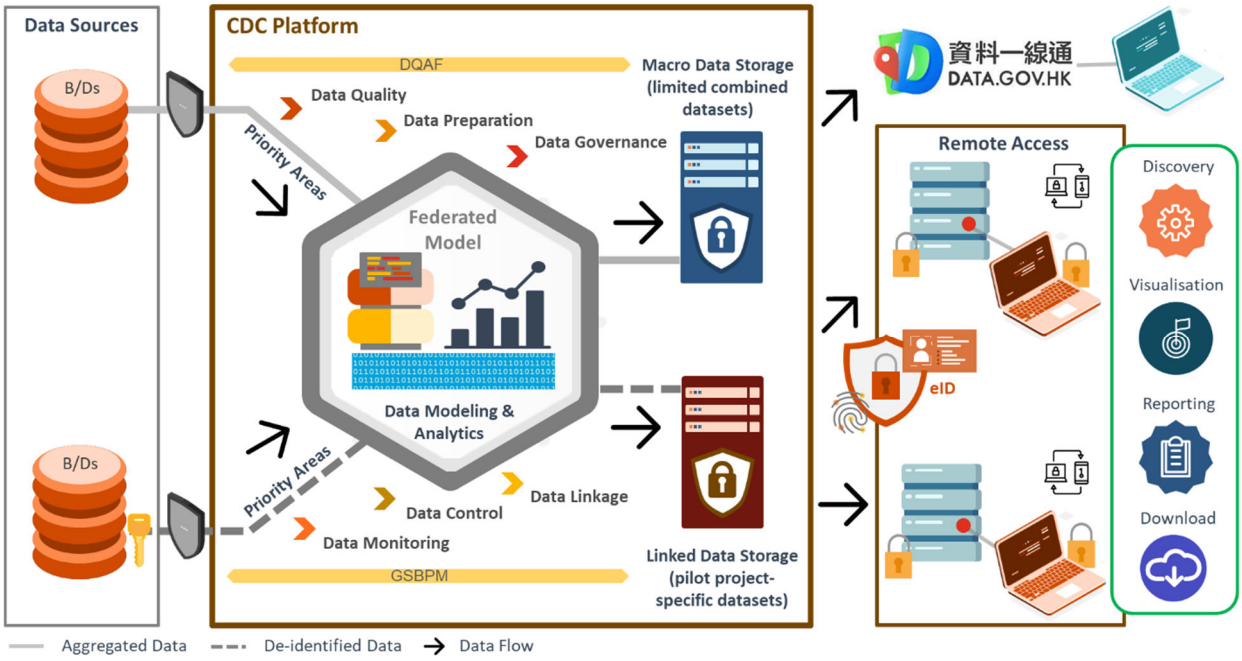
2.1.3.1 IT Framework

A practical IT framework for CDC implementation is crucial in facilitating the contribution of data from various Government B/Ds, NGOs and other public/private organisations. It should also enable collaboration and provide flexibility to cater for future extension or expansion. An IT framework has thus been designed to support the federated model for CDC development by minimising any disruption to various data sources caused by technical integration, and at the same time, respecting the ownership, control and confidentiality of corresponding data sources.

2.1.3.1.1 IT Framework for Foundational Mode

As illustrated in **Figure 9**, the proposed technical architecture under this IT framework for the Foundational Mode would support the federated model in a two-tiered structure for macro-level and micro-level data from B/Ds for the processing and dissemination of children related data. Data shared by B/Ds, whether at macro-level or micro-level, are transmitted through a secure conduit over the Government Backbone Network ("GNET") with encrypted Application Programming Interface ("API") through a file interchange mechanism to CDC. In order to cater for the security, flexibility and volume of data to be processed, the IT framework for CDC should leverage the Government Cloud Infrastructure Services (GCIS) for (i) higher data, application and system security; (ii) better computing resource scalability; and (iii) more flexibility for on-demand service provision when necessary.

Figure 9 Proposed Technical Architecture under the IT Framework for Foundational Mode



At the macro-level, the proposed technical architecture of this IT framework would allow non-identifiable, aggregated data shared by contributing B/Ds to be stored centrally in the macro data storage repository of CDC. Additional data modelling and processing of these aggregated data to produce value-added thematic-based datasets of interests to the public would be supported through data management and analytics software tools, including commercial software, such as SAS or SPSS, or opensource tools, R and Python. The adoption of DQAF and statistical tools would provide support for the necessary data quality, preparation and governance functionality under this IT framework. The resulting thematic-based, non-identifiable datasets would be made available to the general public through the Public Sector Information (PSI) Portal (i.e. DATA.GOV.HK).

At the micro-level, for the proposed pilot projects, contributing B/Ds would extract the identifiable data in their respective databanks then apply the de-identification algorithm provisioned by CDC Data Governance Unit with a one-time encryption key for de-identifying the sensitive personal data in the dataset for the purpose of the specific pilot project before passing the data to CDC. There are a few proven de-identification algorithms, such as Chorus and Google Differential Privacy Library.[20]. These algorithms are well accepted by research and statistics communities around the world and could be considered for adoption by CDC.

The transmitted de-identified data would then undergo data linkage processes by the Data Analytics and Linkage Unit of the CDC Project Office, following which the processed resulting sample datasets would be stored in the Linked Data Storage repository of CDC. The adoption of GSBPM, DDI, DQAF and statistical tools would ensure appropriate data monitoring and control measures are applied to the data linkage process, while supported by the necessary data quality, preparation, and governance functionality under this IT framework.

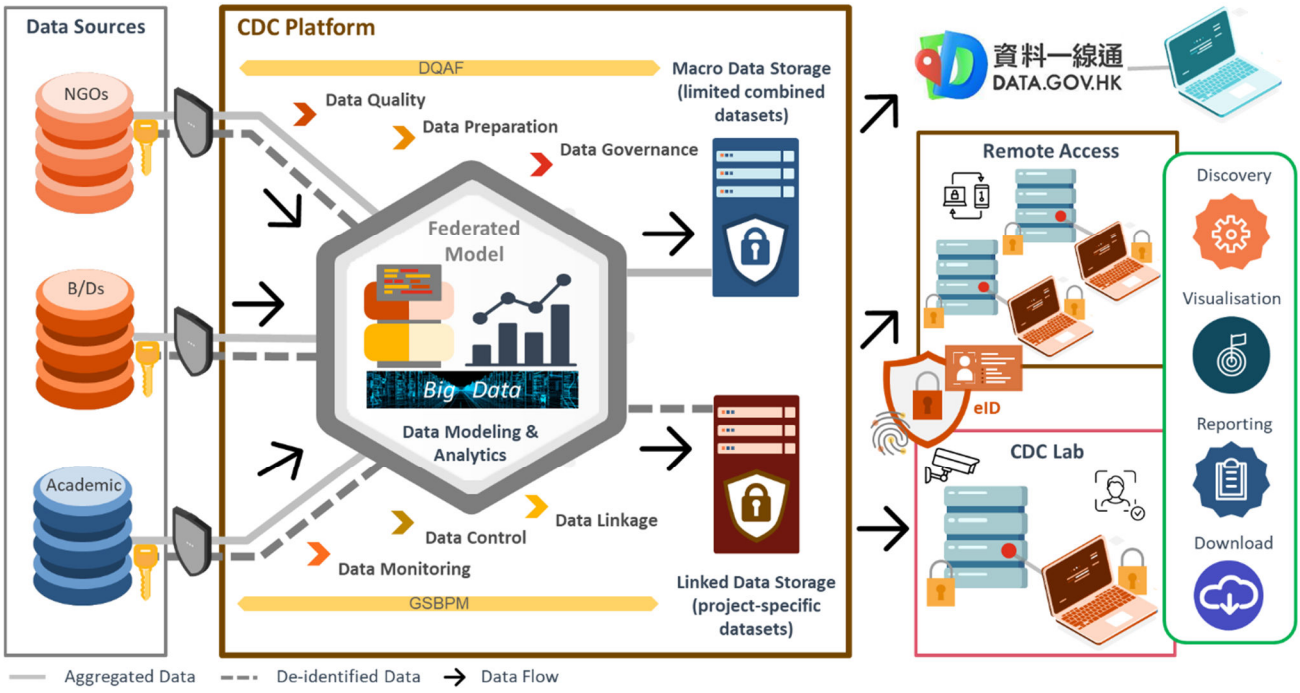
The resulting sample dataset would be made available for the duration of the pilot projects for access by research teams of corresponding pilot projects through a designated sandbox in virtual computing environment (i.e. secure virtual machines or VMs) hosted at the Government Cloud Infrastructure. It is recommended that the Hospital Authority Data Collaboration Laboratory (HADCL) Self-service Data Platform should be referenced for the remote access with strengthening security measures, such as support of electronic identity (eID) and two-factor authentication for safe guiding purposes. End user data analytics tools for drafting research hypotheses or proposals and conducting research on the available dataset would be provisioned through the sandbox VMs.

[20] https://www.nist.gov/itl/applied-cybersecurity/privacy-engineering/collaboration-space/focus-areas/de-id/tools

2.1.3.1.2. IT Framework for Enhancement Mode

For the Enhancement Mode, the two-tiered, federated model IT framework is designed to ensure sensitive data to be remained within relevant individual databases for control and ownership. Under the proposed technical architecture of CDC, sharing of micro-level data would be conducted only through an agreed data sharing mechanism initiated by CDC based on requests from CDC authorised users, and with agreement from corresponding data source owners, i.e. B/Ds, NGOs and other external organisations, such as academic. Upon request and after internal clearance, individual data source owners would extract relevant data and share the de-identified datasets with CDC through the secure conduit over GNET with encrypted API of the file interchange mechanism to CDC. **Figure 10** illustrates the proposed technical architecture of the IT framework for Enhancement Mode, which is a full-featured design that is capable of supporting the production of Enhancement Mode and beyond for both macro- and micro-level data processing and analysis of different research projects relating to the various dimensions of children welfare.

Figure 10 Proposed Technical Architecture under the IT Framework for Enhancement Mode



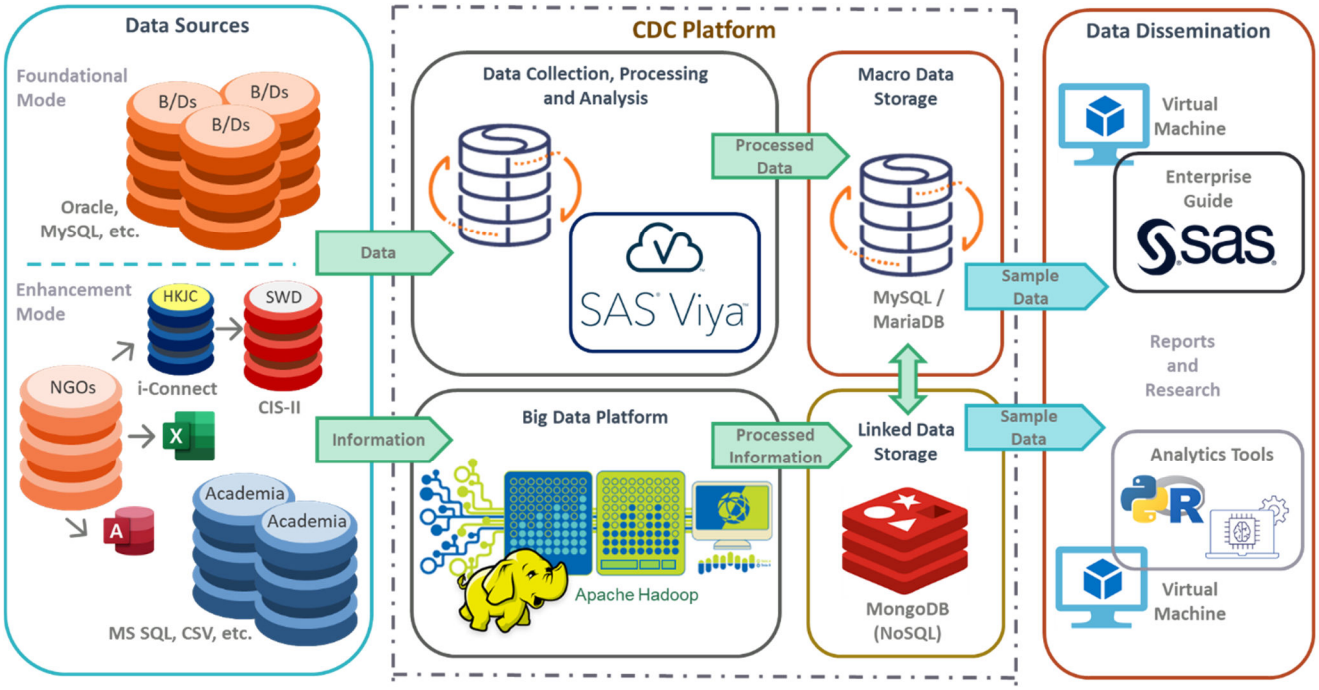
Under the Enhancement Mode, additional data analytic capability would be introduced through the adoption of the Government's Big Data Analytics platform. The existing data processing, data quality management and data governance functions for both macro- and micro-level data would still be applied. The de-identification algorithm and data linkage mechanism with further fine-tuning based on lesson learnt from the two pilot projects would continue to be enforced when sensitive personal information are being handled.

One additional feature that the proposed technical architecture of the IT framework under the Enhancement Mode could enable is the gradual introduction of the CDC Lab. Modelled after the similar facility of the HADCL, the CDC Lab would allow authorised

users of CDC to conduct research of sensitive de-identified data loaded on designed sandbox VMs. To access the CDC Lab, the individuals must be pre-registered, pre-approved, and authorised users of CDC.

The corresponding software components supporting both the Foundational Mode and Enhancement Mode of CDC are illustrated in **Figure 11**. The proposed software list consists latest proven solutions and tools for data analysis, such as SAS Viya, commonly adopted by National Statistical Offices, including the Census and Statistics Department of the Hong Kong Government (See **Appendix G** for a list of proposed software for data analysis). With the exception of the Big Data Platform (Apache Hadoop) which would provide additional functionality to support the Enhancement Mode, all other software components are essential to the rollout of the Foundational Mode.

Figure 11 Software Components Supporting the Foundational and Enhancement Modes

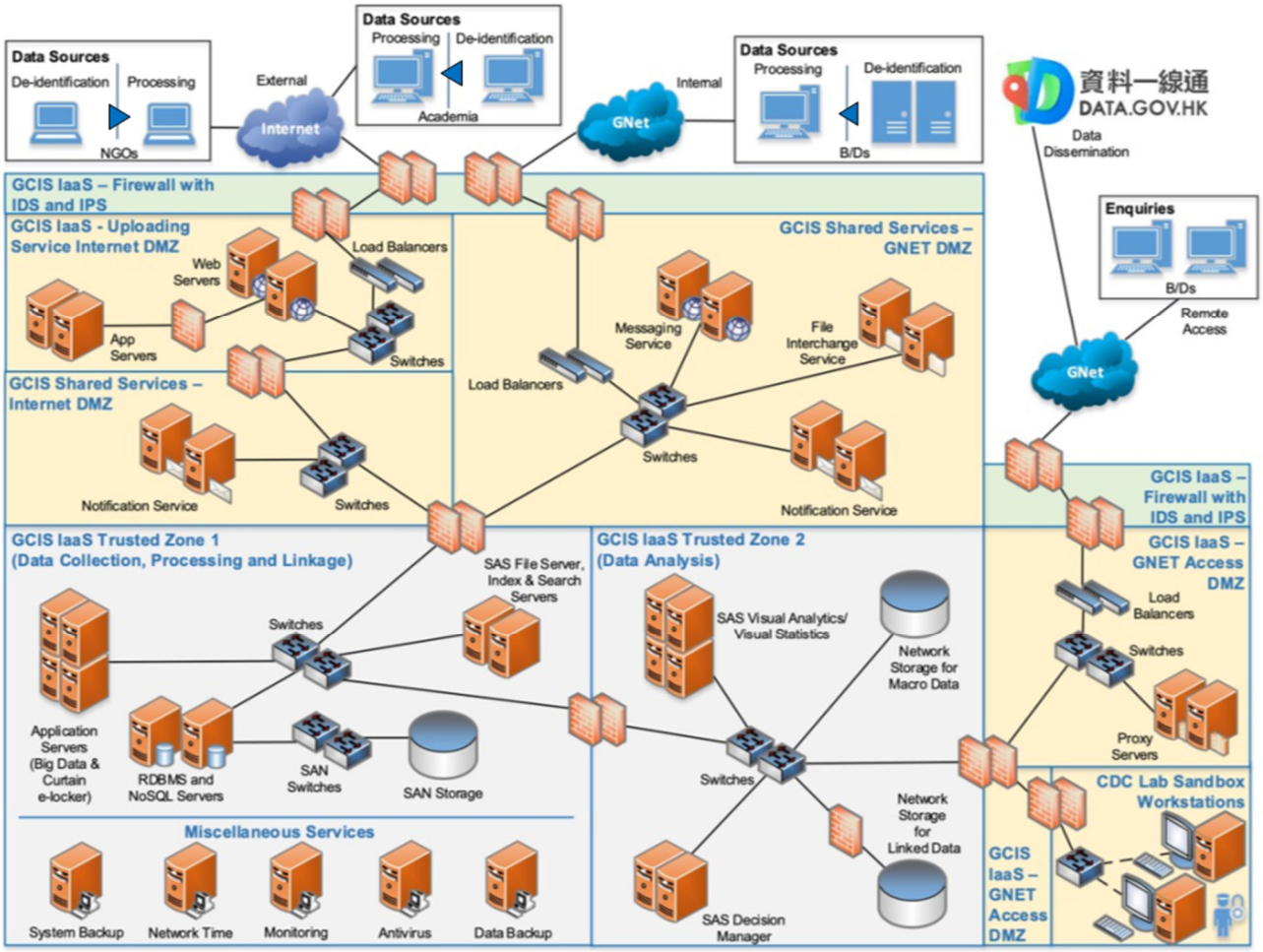


Data exchange of the CDC takes place at data collection and data dissemination, while data processing and analysis are conducted within the CDC platform. From the data exchange planning perspective, CDC would initially facilitate the contribution/collection of relevant data under the Foundational Model for the identified priority areas from selected B/Ds, including EDB, DH, SWD, HKPF and HA. Data contributed/collected would be through pre-defined data codebooks and extracted from the corresponding databases of the corresponding systems of relevant B/Ds, such as Oracle or MySQL. While preparation and alignment of the extracted data would be carried out by the corresponding B/Ds individually, de-identification would be conducted individually by B/Ds and supported centrally by the Data Analytics and Linkage Unit of CDC. Frequency of data collection/contribution would be aligned among relevant B/Ds according to the specifications of the research projects (See **Section 2.1.2.3** for the proposed data alignment plan).

Under the Enhancement Mode, additional data sources from NGOs and the academia is anticipated to be incorporated in CDC. The extent to which data could be exchanged with CDC will depend on various factors such as the purpose of data exchange, technical maturity of databases, quality of data and measures to handle privacy concerns (e.g. consent from beneficiaries or research subjects) among others. For NGOs, depending on the technological maturity of the organisations, the data exchange mechanism of CDC could support direct import of data files in MS Excel, desktop database solutions such as MS Access, or data uploaded via i-Connect of Hong Kong Jockey Club (HKJC) to SWD's Client Information System II. Similarly, different data exchange protocols are available for academia including direct data extraction from RDMS, such as MS SQL Server or in plain text CSV format exported from SPSS, depending on the tools being utilised by academia as well. Again, the frequency of data collection/contribution would depend on the specifications of the corresponding research projects. Data dissemination from CDC, regardless of macro data for trend monitoring or linked de-identified data, would be made available to B/Ds and researchers based the research topics as requested. Software package for statistical analysis or analytic tools such as SAS Enterprise Guide, SPSS, R, Python or machine learning algorithm could be leveraged for reporting and researching purposes.

The high-level physical design diagram at **Figure 12** on the next page illustrates the major components and elements of CDC under the proposed technical architecture covering both the Foundational Mode and Enhancement Mode. This high-level physical design provides an all-encompassing architectural view of CDC being built under GCIS leveraging Infrastructure-as-a Service (IaaS) as the foundation for optimal resource utilisation and the necessary security required while other shared services, such as File Interchange Service, Messaging Service and Notification Service for data exchange and communication with B/Ds via GNET, will also be subscribed for supporting application services as well, while Notification Service and servers for uploading services are established for data exchange and communication with external parties, including NGOs and Academia, via the internet. For security and data integrity of the data being collected, processed and analysed in CDC, different network zoning for CDC is proposed, including two DMZ zones channelled for data submission from GNET and the internet. Similarly, for data security purposes, B/Ds, NGOs and Academia are advised to establish the necessary logical partitioning between data sourcing and data de-identification processes as illustrated in the diagram for data collection by the CDC. Tandem trust zones, one for data collection and processing, and one for data analysis, segregated by firewall are proposed to ensure the data confidentiality and integrity are maintained for data in transit and at rest within CDC. Other DMZ zones are setup for data dissemination via GNET for remote access by B/Ds or from the proposed CDC Lab. The network zoning and security measures should be revisited upon the implementations of the Foundational Mode and the Enhancement Mode of the CDC respectively to ensure the setup and configuration align with the latest security requirements of the Government.

Figure 12 High-level Physical Design Diagram



2.1.3.2 Data Exchange with Third Parties

2.1.3.2.1 IT Architecture for Data Exchange

Apart from the consolidated data from B/Ds, data consolidation can be extended to third parties as when the CDC platform has been well developed and the preparedness of the third parties for the contributions of the collected data.

The subscribed Shared File Interchange Service from GCIS would facilitate as the universal data exchange mechanism between CDC and all external data sources, whether they are government B/Ds or external NGOs. The Shared File Interchange Service of GCIS would allow CDC to collect from various data sources from Government B/Ds via GNET, whereas a separate uploading services would be setup for collecting from data from NGOs and Academia via the internet, regardless of the formats of survey data and statistical returns being transferred from the data sources to CDC, i.e. as pre-processed data files or direct data exchange files from systems, or the technical maturity and transmit protocols of the data source organisations.

In general, when selecting an external party to conduct data exchange with, CDC may wish to give consideration to the following criteria:

(i) purpose of data exchange;

(ii) technical maturity of the database;

(iii) quality of data (e.g the frequency of data update and accuracy of data); and

(iv) anticipated benefits for the public against the cost for exchange (including privacy and ethical implications)

Based on information collected from earlier stakeholder consultation exercises, the key features of several databases on children data owned by NGOs, academia, and a HKJC Charity Trust project are summarised at **Table 22**. Varying level of technical maturity could be observed, with HKJC and HKU likely to be more ready for such exchange when compared with the other two organisations. Previous stakeholder engagement also suggests that NGOs with lower level of technical maturity tends to have more concern over the liability of sharing inaccurate data. For a summary of the current state of these databases, please refer to **Appendix H.**

Table 22 Overview of Relevant Databases

Database Name	The Hong Kong Child Development Indicators	KeySteps@JC	Young Children's Development Indicators of Hong Kong	1997 Birth Cohort Study
Database Owner	Boys and Girls Club Association of Hong Kong	Hong Kong Jockey Club	Hong Kong Society for the Protection of Children	School of Public Health, LKS Faculty of Medicine, The University of Hong Kong
Brief Description	The Hong Kong Child Development Indicators (CDI) has been in use since established in 2006. It aims to: i. Establish a comprehensive database of children's information. ii. Have a clearer grasp of the living conditions and development trends of children in Hong Kong and indicate the future	KeySteps@JC has been in use since 2018. It is adopting a cloud approach for data management so that multi-disciplinary teams and external stakeholders will draw upon one unified dataset. It aims to: i. Provide holistic support (health, emotion, social skills, cognitive and language development) to build stronger foundations for better life outcomes of the children;	The Young Children's Development Indicators of Hong Kong (YCDI) has been in use since/ established in 2014. It aims to: i. Provide reliable and periodical macro data for young children professionals, policy makers, NGOs, and other people concerned about young children's wellbeing; and ii. Facilitate higher quality of child services and children policy.	1997 Birth Cohort Study has been implemented since 1997. It aims to: i.To explore determinants of health (e.g. genetics, behavioural, anthropometrics, and biomarkers) in a non-Western setting, which may shed light in the understanding of drivers of diseases and hence identifies potential targets of intervention; and i.In the long run, the "Children of 1997" birth cohort shall serve as a unique epidemiologic resource available to bona fide researchers to carry out their research, following the framework of ALSPAC and UK Biobank.

Database Name	The Hong Kong Child Development Indicators	KeySteps@JC	Young Children's Development Indicators of Hong Kong	1997 Birth Cohort Study
	development direction of policies and social services. iii. Effectively monitor Hong Kong's children's policies and services; whether Hong Kong is a ""child-friendly city"". iv. Set goals for children's policies and social services. v. Provides valid and practical suggestions to the Government and society.	ii. Enhance parent-child attachment and promote parent/grandparent/caregiver well-being to strengthen family functioning; iii. Enhance capacity building, enrich learning environment and school curriculum to create quality learning environment; iv. Develop a child-focused community support model so to create community space to foster social connectedness; v. Develop the first cross-disciplinary one-stop web		

Database Name	The Hong Kong Child Development Indicators	KeySteps@JC	Young Children's Development Indicators of Hong Kong	1997 Birth Cohort Study
		portal and child databank, as well as use evidence-based research to inform future policy design and serve as a launchpad for big data foundations in medical, welfare and education; and vi. Inform programme design and refinement, intervention, enhance programme effectiveness.		

Database Name	The Hong Kong Child Development Indicators	KeySteps@JC	Young Children's Development Indicators of Hong Kong	1997 Birth Cohort Study
Example(s) of data	• Demographic characteristics • Family background • Education status • Health and safety • Social behaviour • Economic Security • Social-environmental factors	• Personal identifier (e.g. Name, etc.); • Parent information (e.g. Name, etc.); • Medical Information; • Social behaviour; and Education status	• Demographic and social characteristics (e.g. age, ethnicity, sex, etc); • Educational characteristics (e.g. School attendance, special education needs, etc); • Economic characteristics (e.g. economic activity status, monthly income, poverty, etc); • Medical characteristics (e.g. death, cancer, body weight, etc.); and • Social Welfare Assistance (e.g. abuse cases, financial assistance, etc.).	• Biospecimens (e.g. hair, blood, saliva, etc.) • Living habit (e.g. sleeping habit, household smoking habit, etc.) • Health measures (e.g. blood pressure, birth weight, mental health, etc.) • Social and economic characteristics (e.g. parental educational attainment, etc.)

Database Name	The Hong Kong Child Development Indicators	KeySteps@JC	Young Children's Development Indicators of Hong Kong	1997 Birth Cohort Study
Data Architecture	Data is updated on an ad-hoc basis, with no data sharing arrangement currently.	Adopts a three-tier database management structure, with regular frequency of update. Both individual and aggregate level of data is available. A data dictionary is also available.	Adopts a single-tier database management structure, with regular frequency of update. Aggregate level of data is available.	Adopts a two-tier database management structure, with data updated on a need basis. Individual level of data and a dictionary are available.

2.1.3.2.2 Data Exchange Work Plan

The key stages and tasks involved in conducting data exchange as well as the parties involved are set out in **Table 23**. When approving the application of third parties for data exchange, the CDC Project Office is suggested to consider the following criteria:

- Having personal identifier (i.e. HKID) as part of collected data
- Explicit consent of data subjects for sharing with CDC
- Regularity of collection of research data

Table 23 Four stages of data exchange with third parties

Stages	Steps	Actions	Parties Involved
Stage 1 – Preparation	Step 1	Prepare the guidelines for dataset and codebook compilation, in format such as CSV and TSV.	CDC Project Office
	Step 2	Submit the details of the projects for the review and approval for data exchange in the CDC platform by the CDC project office.	Third Parties
	Step 3	Review and approves third parties' projects for data exchange in the CDC platform	CDC Project Office
Stage 2 – Data Compilation and Conversion	Step 1	Prepare the introduction of the projects, codebook and de-identified dataset according to the guidelines issued by the CDC project office.	Third Parties
	Step 2	Conduct data encoding as no personal identities should be stored in the dataset to be exchanged	Third Parties
Stage 3 – Data Checking and Uploading	Step 1	Check and approve the codebook and de-identified dataset submitted by the third parties.	CDC Project Office
	Step 2	Upload the codebook and de-identified dataset after the approval by the CDC project office	Third Parties
Stage 4 – Data Exchange	Step 1	Retrieve the codebook and de-identified dataset of various. The data exchange will be conducted in the CDC platform.	Third Parties

2.1.3.3 Estimated Size of CDC Database

2.1.3.3.1 Foundational Mode

Based on the initial codebook information provided by the five B/Ds, namely: Education Bureau (EDB), Hospital Authority (HA), Social Welfare Department (SWD), Hong Kong Police Force (HKPF), Department of Health (DH) and Hospital Authority (HA), on the identified priority areas covering "Children with Risk of Abuse and Neglect" and "Children with SEN", the aggregated data size according to the data type definitions is estimated to be 0.50 MB annually. Given the fact that there are approximately 2,564,705 entries for the aggregated data based on the codebooks

reported by these five B/Ds[21] and HA, the total combined data size is approximately 69.3 GB annually. Since it is necessary to maintain children data from infancy to adulthood to facilitate potential longitudinal studies, the data would need to maintain for 25 years (i.e. 18 years plus another seven years according to the standard record disposal schedule of the Government). Thus, the estimated sizing of CDC under the Foundational Mode for aggregate data from contributing B/Ds as of 2022 is around 3.38 terabytes (TB) based on equal coverage of data in (i) Health and Safety, and (ii) Education and Skills of the children well-being dimensions. Assuming an annual growth rate of 15%, and the adoption of the data storage ratios for data collection, processing and dissemination as illustrated in **Table 24** below, a total of **11.9 TB** storage is required.

Table 24 Estimated Sizing of CDC under the Foundational Mode

Process	Data Storage Ratio	Data Sizing Requirements
Data Collection	1	3.4 TB
Data Processing	2	6.8 TB
Data Dissemination	0.5	1.7 TB
	Total:	11.9 TB

Applying the 11.9 TB as the baseline in 2022, projecting over five years starting from year 2023 (assuming the duration of the Phase 1, the Foundational Mode, and subject to further revision), the storage requirements for CDC under Phase 1 is as follows:

Table 25 Projected Storage Requirements for CDC under Foundational Mode

Projected CDC Data Sizing for Phase 1 – Foundational Mode					
Year	2023	2024	2025	2026	2027
Size (TB)	13.6	15.7	18.0	20.7	23.8

A resilient setup comprising server configuration of the SAS Viya as well as other servers and shared services requirements proposed under is illustrated:

[21] The number of entries for the aggregated data is based on the number of entries per year of relevant databases as reported by the five B/Ds. This number will be further validated with B/Ds.

Table 26 Server Components and Shared Services under Foundational Mode

Server Components	Quantity
1. SAS Viya (CAS Worker Notes + CAS Controller Node)	7
2. Curtain e-Locker Server for SAS files	2
3. Maria/MySQL Database Servers	2
4. MongoDB (NoSQL) Servers	2
5. File Servers	2

Shared Services (for GNET DMZ only)
6. Messaging Service
7. Notification Service
8. File Interface Service

2.1.3.3.2 Enhancement Mode

For Phase 2, the Enhancement Mode, assuming to commence in 2028 for another duration of 20 years to fulfil the requirements of maintaining the children data for overall period of 25 years, the CDC data size is deduced from the same codebooks received from the 5 B/Ds and HA but further extrapolated to cover a total of 19 potential data contributing candidate B/Ds and external organisation, such as:

- Government B/Ds: EDB, DH, HA, SWD, HKPF, Home Affair Department, Housing Authority, C&SD, Department of Justice and Juvenile Court;
- External Organisations: the Boy's and Girl's Clubs Association, Hong Kong Society for the Protection of Children, Hong Kong Council of Social Services, and Hong Kong Jockey Club;
- Academia: School of Public Health (University of Hong Kong), Department of Paediatrics (the Chinese University of Hong Kong), Faculty of Law (University of Hong Kong), Faculty of Law (the Chinese University of Hong Kong) and Advisory Committee on Mental Health.

The projected sizing of CDC under the Enhancement Mode for aggregate data from these identified contributing B/Ds and organisations as of 2022 is approximately 19.2 TB based on equal data coverage for (i) Health and Safety and (ii) Education and Skills, i.e. 1:1 ratio; and (iii) Material Well-being, (iv) Behaviours and Risks, and (v) Family and Peer Relationship, i.e. 1/3: 1/3: 1/3, of the five children well-being dimensions respectively. Assuming an annual growth rate of 15% and the adoption of the data storage ratios for data collection, processing and dissemination as illustrated in **Table 27** below**,** a total of **67.3 TB** storage is required.

Table 27 Estimated Sizing of CDC under the Enhancement Mode

Process	Data Storage Ratio	Data Sizing Requirements
Data Collection	1	19.2 TB
Data Processing	2	38.5 TB
Data Dissemination	0.5	9.6 TB
	Total:	67.3 TB

Applying the estimated 56.3 TB as of 2022 and further projecting over 25 years starting from year 2023, as well as assuming the duration of the Phase 2, the Enhancement Mode, for a period of 20 years immediately after Phase 1, i.e. commencing in 2028), the storage requirements for CDC under Phase 2 is as follows:

Table 28 Projected Storage Requirements for CDC under Enhancement Mode

Projected CDC Data Sizing for Phase 2 – Enhancement Mode										
Year	2028	2029	2030	2031	2032	2033	2034	2035	2036	2037
Size (TB)	156	179	206	237	272	313	360	414	476	548
Year	2038	2039	2040	2041	2042	2043	2044	2045	2046	2047
Size (TB)	630	724	833	958	1,101	1,267	1,457	1,675	1,926	2,215

The following server configuration of the SAS Viya as well as other servers and shared services requirements are proposed under a resilient setup to cater for the above data size of CDC under the Enhancement Mode:

Table 29 Server Components and Shared Services under Foundational Mode

Server Components	Quantity
1. SAS Viya (CAS Worker Notes + CAS Controller Node)	14
2. Big Data Server Platform	6
3. Curtain e-Locker Server for SAS files	4
4. Maria/MySQL Database Servers	4
5. MongoDB (NoSQL) Servers	4
6. File Servers	4
7. Uploading Servers (Web and Applications) for Internet	4

Shared Services (for GNET)
8. Messaging Service
9. Notification Service
10. File Interface Service

Shared Services (for Internet)
11. Notification Service

The approach and assumptions for the sizing estimation is further elaborated at **Appendix I**.

2.1.4 Implementation Dimension

2.1.4.1 Implementation Roadmap

Considering the anticipated scale, complexity and perceived sensitivity of the development of CDC, a phased approach is suggested for CDC implementation. The Foundational Mode of CDC is proposed to be undertaken as a first step given its likelihood of higher public acceptance and lower privacy concerns. This will be followed by the launch of the Enhancement Mode. By adopting a phased approach for implementation, it provides opportunities to identifying rooms for improvement and enhancement throughout the execution. Through regular communications on the development progress and continuous engagement with key stakeholders, this can also enable gradual build-up of public acceptability of and trust in CDC overtime.

To demonstrate benefits of CDC to potential users and the public early on, pilot projects are proposed to be conducted at the initial stages of CDC development to shed light on important child-related issues, while showcasing benefits of CDC to the public. It also offers an opportunity to identify potential challenges to be tackled before full implementation of CDC.

2.1.4.1.1 Pre-requisites and Dependencies

Upon successful implementation of the Foundational Mode of CDC, the CDC Project Office could consider further development of the Enhancement Mode of CDC, the following pre-requisites should be taken into consideration:

Dimension	Pre-requisites to be fulfilled for implementation of Enhancement Mode of CDC
Business and Legal Dimension	• Approval obtained from the oversight body of CDC with regard to the objective and scope of CDC's Enhancement Mode • Consultation with the Privacy Commissioner for Personal Data and engagement with experts (i.e. with expertises in data linkage and de-identification) • Establishment of a fully commissioned data ethics and privacy panel to ensure ethical use of children's data
Data Dimension	• Update of consent form of data providers to ensure future collection of data covers usage of de-identified data for data linkage projects • Data alignment to be performed for topics of concern based on consensus reached with data providers for linkage projects • Informed consent to be sought from data subjects for Enhancement Mode of CDC for data linkage projects and longitudinal studies where applicable
Technology Dimension	• A technically stable platform in place with well-defined operations and procedures

Dimension	Pre-requisites to be fulfilled for implementation of Enhancement Mode of CDC
Implementation Dimension	• Satisfactory completion of PIA • Availability of funding for contracting out data linkage and/or longitudinal studies

2.1.4.1.2 Timeline and Key Tasks

An implementation roadmap was devised based on the four dimensions under the development framework of CDC, namely Business and Legal dimension, Data dimension, Technology dimension and Implementation dimension. It is anticipated that both the Foundational and Enhancement Modes will be implemented over a period of ten years. The Enhancement Mode, which is stated to be implemented from Year 6 onwards, is based on the assumption that the pre-requisites indicated in **Section 2.1.4.1.1** will be fulfilled by Q1 of Year 6. The Enhancement Mode may be implemented at an earlier time subject to periodic assessment of the oversight body of CDC. The pilots for the two priority projects are anticipated to commence from Year 2 onwards following the setting up of the governance and execution function of CDC as well as development of the de-identification mechanism in Year 1.

Upon the completion of the pilot projects, an evaluation of the pilot projects shall be conducted to assess whether the pilots are able to demonstrate to the public the benefits of CDC and identify key challenges in the run up to Enhancement Mode. Subject to the assessment of the WG overseeing CDC and satisfactory outcome of the pilots, subsequent consideration may be given to the commissioning of other research projects to further demonstrate to the potential benefits of CDC.

While the topic, scope, and duration of future research projects are to be decided by the WG overseeing CDC, one of the potential directions of the projects could be thematic studies on key priority areas as steered by the CoC. Such thematic studies may leverage the data collected from well-established research initiatives. For instance, a thematic study on child health development could potentially be conducted leveraging the existing physical fitness data of school children collected by the School Physical Fitness Award Scheme since 1990, with which linkage may be created to understand children's development progress when moving to different schools or districts. This could enable better understanding of health outcomes and fitness achievement of children, both at a point in time and longitudinally, that will support decision-making of policymakers and analysis of researchers of patterns in the distribution of fitness level across the population.

Key tasks identified under each dimension are illustrated below.

Table 30 Key Tasks to Be Completed by Phases

Key Tasks to Be Completed		Phases		
		Year 1 - 5 (Foundational Mode)	Year 6 - 10 (Enhancement Mode)	Periodically
Business and Legal Dimension				
1.	**Project Initiation and Governance Setup**			
1.1	Agree on framework for CDC development and secure resources (For Foundational and Enhancement Modes)	√	√	
1.2	Set up governance structure for CDC	√	√	
2.	**Consultation and Engagement**			
2.1	Identify key stakeholders (internal and external) for briefing on commencement of CDC development	√		
2.2	Engage stakeholders for system analysis and design (SA&D) of CDC development	√	√	
2.3	Engage key stakeholder groups for feedback on CDC development on a periodic basis[1]		√	√
2.4	Engage stakeholders at the commencement of pilot studies, progress updates and release of findings	√		
Data Dimension				
3.	**Data Governance and Alignment**			
3.1	Consult with the Privacy Commissioner on data privacy and ethical issues	√		
3.2	Establish data governance mechanism with consideration of consultation of DOJ and PCPD	√		
3.3	Carry out data alignment exercise for children data for priority areas (e.g. key demographic data)[3]	√		
Technology Dimension				
4.	**Development of Foundational Mode**			
4.1	Engage vendor for development of Foundational Mode of CDC	√		
4.2	Conduct SA&D for CDC development (Foundational Mode)	√		
4.3	Develop and launch CDC (Foundational Mode)	√		

Key Tasks to Be Completed		Phases		
		Year 1 - 5 (Foundational Mode)	Year 6 - 10 (Enhancement Mode)	Periodically
4.4	Operate CDC (Foundational Mode) with trend monitoring function[2]	√	√	
5.	**Development of Enhancement Mode**			
5.1	Engage vendor for development of Enhancement Mode of CDC		√	
5.2	Conduct SA&D for CDC development (Enhancement mode)		√	
5.3	Develop and launch CDC (Enhancement Mode)		√	
5.4	Commission and facilitate data linkage projects/ longitudinal studies under the Enhancement Mode subject to approval from CDC governance[2]		√	
Implementation Dimension				
6.	**Launch of Pilot Projects**			
6.1	Set up interim governance functions for pilot projects (e.g. ethics function and dedicated project teams etc.)	√		
6.2	Develop de-identification mechanism as part of executive function	√		
6.3	Confirm objectives and commission pilot study upon consultation with DOJ and PCPD on de-identification mechanism	√		
6.4	Conduct pilot study - coordination and data alignment for pilots in two priority areas[3]	√		
6.5	Conduct pilot study - Data analysis and reporting	√		
6.6	Evaluation of pilots and consider projects for further execution (e.g. updating existing project and/or commissioning of new projects) [4]	√		
7.	**Periodic Review and Assessment**			
7.1	Publish regular reports on the development progress of CDC			√
7.2	Conduct on-going review and monitoring for CDC (e.g. changes in key priority areas, areas for improvement, system audit)			√
8.	**Promotion and Publicity**			
8.1	Promote the key role and benefits of CDC (e.g. engage public figure, official videos, social media content. thematic reports)	√	√	

Note 1: Including the engagement of stakeholders during the various phases of system design, such as User Acceptance Test and softlaunch.

Note 2: Both Tasks 4.4 and 5.4 are subject to regular review under Task 7.1.

Note 3: Tasks 3.3 and 6.4 are to be conducted concurrently. Key stakeholders should also be further engaged as part of Task 2.1.

Note 4: The commissioning of further research projects are subject to the assessment of the governance body taking into account the outcome and evaluation of pilots, with the topic an scope to be decided by the WG overseeing CDC.

2.1.4.2 Potential Pilot Projects on "Children with Risk of Abuse and Neglect" and "Children with SEN"

2.1.4.2.1 Children with Risk of Abuse and Neglect

Research Questions

To facilitate the future scoping of the pilot studies, a list of potential research questions (RQ) for "Children with Risk of Abuse and Neglect" has been developed for illustration below. It is to note that the list of RQ is not meant to be exhaustive and is solely for the purpose of illustrating potential benefits of data linkages only.

Table 31 Illustrative RQ for Children with Risk of Abuse and Neglect

Research Questions (RQ)
RQ1 – What are the macro-trend and incidence rates of reported child abuse cases in Hong Kong?
Rationales: • To present the incidence rates of reported child abuse cases • To illustrate the macro-trend of reported child abuse cases across multiple years, and to monitor effectiveness of service provision and policy intervention Statistics and tables proposed: • Types of abuse (e.g. physical abuse, child neglect, psychological abuse and sexual abuse) • Characteristics of victims, abusers, non-abusive parents of reported cases of child abuse • Yearly trend of reported abuse cases
RQ2 – What are the underlying differences for case reporting across databases from SWD, HA and HKPF?
Rationales: • To compare reported child abuse cases across various databases from SWD, HA and HKPF • To explore the differences in reporting Statistics and tables proposed: • Comparison between data of reported child abuse cases in databases of SWD, HA and HKPF • Types of Abuses
RQ3 – What are the risk factors of reported child abuse cases?
Rationales: • To identify the demographic backgrounds, health indicators and family-related factors of reported child abuse cases • To enable early identification of potential child abuse cases and risk assessments Statistics and tables proposed:

Research Questions (RQ)
• Demographic backgrounds of reported child abuse cases (e.g. gender, age, educational attainment, school non-attendance and academic results) • Health-related factors (e.g. A&E utilisation and attendance rates, A&E diagnosis and medical records, mental and physical health status, growth, immunisation records, developmental problems) • Family-related factors (e.g. family income, poverty, type of housing and socio-economic status of the parents, teenage pregnancy, mother/parent with mental illness, substance abuse, domestic violence) • Comparison of demographic, health and family risk factors between reported child abuse and non-child abuse cases
RQ4 – Are there any geographical differences in distribution of reported child abuse cases?
Rationales: • To illustrate the geographical distribution of reported child abuse cases • To explore locational differences by mapping reported child abuse cases and service provision • To identify service gaps for specific district and identify improvement areas in policy design Statistics and tables proposed: • Reported child abuse cases by districts (i.e. District Council (DC) districts and Constituency Areas (CA)) • Association between district-based services and reported child abuse cases • Types of abuse by districts

Expected Outcomes

Specifically, based on the illustrative research objectives and proposed statistics, four expected outcomes could be identified, as follows:

RQ1

Child abuse and neglect has attracted considerable attention in recent years because of its increasing trend. In response to RQ1, the incidence rates of all types of abuse could be illustrated in a time series manner. The findings could help to monitor and review the performance of relevant services, intervention programmes and/ or policies, strengthen relevant intervention programmes, and design and implement new services where appropriate.

RQ2

The matched cases represent the abused cases appeared in the three databases. Supplemented by the other linked databases, the data fields will be rich and provide a comprehensive analysis of the child abuse reported case, specifically for trend monitoring and risk profile analysis. In response to RQ2, the findings of reporting gaps

could help to enhance existing reporting mechanism and strengthen the coordination between B/Ds and improve relevant services.

Three possible reporting gaps can be explored:

- **Cases found in HA but not in the Child Protection Registry (CPR) of SWD imply possible missed investigation by Multi-disciplinary Case Conference (MDCC) of SWD:** Children presented to clinical services and being identified by medical professionals as abuse under the HA CMS are likely to be physically injured, sexually abused, at risk of neglect, etc. Normally, these patients should be referred to MDCC for further investigations. The possible reasons of not showing in SWD's CPR could be (i) not abused cases as judged by MDCC, but probably presenting welfare needs; or (ii) missed investigation by MDCC. Analysis of the profile of these missed cases could provide empirical information to identify which subgroups of suspected abused children may be classified as welfare cases and to test the hypothesis if non-injured abused children are more likely to be missed in intervention.
- **Cases found in HKPF record but not in SWD's CPR imply possible missed investigation by MDCC of SWD:** Children presented to police record and being identified as abuse are likely with criminal nature. Supposedly, these cases should be referred to MDCC for psycho-social investigation and intervention. If there are cases in this category, missed investigation by MDCC may happen. Analysis of the profile of these missed cases could provide empirical information to understand the welfare needs of these cases.
- **Cases found in SWD's CPR only imply no medical needs or missed diagnosis in HA as well as cases not entitled for arrest for prosecution:** Similarly, abused cases recorded in SWD's CPR only could be compared to the reference group (Outcome 1) on the risk profile. It could provide information regarding the profile and factors that the abused cases did not appear in the HA system (possible cases without medical needs or missed diagnosis in HA) or police record (abused cases, not entitled for arrest or prosecution).

RQ3

Risk model analysis could be performed by using non-abuse cases obtained from the (i) DH – CASIS; (ii) HA – CMS / clinical database; and/ or (iii) EDB – STIMS. In response to RQ3, the abused cases could be compared to ordinary children received services from DH and HA, or ordinary students in the education system. The findings would help identify child-related, family-related, health-related and psycho-social risk factors of child abuse.

RQ4

The geographic distribution of child abuse reported cases could be illustrated with mapping the related district-based services. In response to RQ4, the findings of geographic distribution could help examine the service coverage in specific district, strengthen the coordination of services across districts and examine the needs of new services.

In relation to the priority area of children with risk of neglect and abuse, a case illustration concerning data linkage study on child abuse is at **Appendix I**.

2.1.4.2.2 Children with SEN

Research Questions

Likewise, a list of potential RQ for "Children with SEN" are as follows[22]:

Table 32 Illustrative RQ for Children with SEN

Research Questions (RQ)
RQ1 – What is the yearly number of children with SEN in Hong Kong?
Rationales: • To present the number of children with SEN • To illustrate the macro-trend across years to monitor effectiveness of service provision and policy intervention Statistics and tables proposed: • Number of children with SEN from databases • Types of SEN • Characteristics of the children with SEN (e.g. school level, types of schools) • Year trend of the number of children with SEN
RQ2 – What are the factors contributing to differences in reporting of children with SEN conditions across various databases?
Rationales: • To compare reported children with SEN across various databases from different B/Ds (i.e. EDB, SWD and DH/ HA) Statistics and tables proposed: • Comparison of number of children with SEN across databases from different B/Ds • Types of SEN
RQ3 – What are the characteristics of children with SEN that can influence health, safety and/ or education outcomes?
Rationales: • To identify the demographic backgrounds, health indicators and family-related factors of children with SEN

[22] The objectives listed are not exhaustive. For illustration only.

Research Questions (RQ)
• To enable early identification of characteristics of children with SEN that may be more inclined to face risks of health, safety and education outcomes and thus facilitate timely intervention Statistics and tables proposed: • Demographic backgrounds of children with SEN (e.g., gender, age, educational attainment, non-attendance, and academic results) • Health-related data of children with SEN (e.g., A&E utilisation and attendance rates, A&E diagnosis and medical records, mental and physical health status, history of being abuse) • Family related data of children with SEN (e.g., family income, poverty, type of housing and socio-economic status of the parents, family history of physical/ mental health issues, substance abuse) • Comparison of demographic, health and family factors between children with SEN and those without SEN • Prevalence of child abuse among children with SEN
RQ4 – What is the prevalent mode and utilisation of health, educational and social service provision (i.e. across or within DC/ CA) for children with SEN?
Rationales: • To illustrate the geographical distribution of children with SEN • To explore locational differences by mapping children with SEN and service provision • To identify service gaps for specific district and identify improvement areas in policy design • To compare the differences in health, safety and/ or education outcome based on different modes of service provision (i.e. across vs. within DC/ CA) Statistics and tables proposed: • The number of children with SEN by both school and residential districts (i.e. District Council districts (DC) and Constituency Areas (CA)) • Utilisation information (e.g., the age when child is provided with service, the types of health/ rehabilitative service received, duration of service received, any duplication of services from different service providers, etc.) • Association between district-based services and children with SEN • Types of SEN by district

Expected Outcomes

Specifically, based on the illustrative research objectives and proposed statistics, four expected outcomes could be identified, as follows:

RQ1

With the increasing number of children with SEN studying in mainstream schools, many parents and stakeholders expressed grave concern on the services and support

on the children with SEN and their caregivers. In response to RQ1, the number of children with SEN for all types could be illustrated across years. The findings could help to:

- Monitor and review relevant services and existing policies;
- Strengthen relevant intervention and programmes; and
- Design and implement new services where appropriate.

RQ2

The matched cases represent the children with SEN recorded in the various databases. Supplemented by the other linked databases, the data fields will be rich and provide a comprehensive analysis of the children with SEN, specifically for trend monitoring and risk profile analysis. On the other hand, the findings of reporting gaps could help identify potential time differences in reporting of children with SEN conditions across databases from different B/Ds. Whilst understanding that differences in reporting will inevitably occur as individual B/Ds would have collected data on children with SEN for their specific operational purpose, such findings could facilitate analysis of the reasons for the time differences, hence strengthening the coordination between B/Ds and improving relevant services.

RQ3

Risk model analysis could be performed by comparison between the characteristics of children with SEN and those without SEN. In response to RQ3, the children with SEN could be compared to the children without SEN on their demographic profiles as well as family and health background. The findings could help to identify child-related, family-related, health-related and psycho-social factors that may influence selected health safety and education outcomes.

RQ4

The geographic distribution of the children with SEN could be illustrated by mapping the related district-based services. In response to RQ4, the findings of geographic distribution could help to:

- Examine the prevalent mode of service provision and service utilisation in terms of geographic distribution;
- Understand whether the mode of service provision and service utilisation has implications on health safety and/ or education outcomes; and
- Enhance future resource allocation and strengthen coordination of services across relevant B/Ds.

2.1.4.2.3 Timeline and Key Tasks for Pilot Implementation

The duration of the pilots for the two priority areas is estimated to take around 18 months upon commissioning of the pilot projects [23]. As the experience of the first pilot project will benefit the implementation of the second one by avoiding possible abortive work, it is proposed that the two pilot projects should be taken forward sequentially. Hence, it will take about 36 months to complete the two pilot projects.

Regarding the implementation, two major stages are proposed (i.e. data matching and data analysis and reporting) as illustrated in **Table 33** below. These tasks are also illustrated at Tasks 6.4 and 6.5 in **Table 30** of **Section 2.1.4.1.2**.

Table 33 Implementation Timeline and Key Tasks for the Illustrations

Stage	Key Tasks	Estimated Duration
Stage 1 **Inception and Data Matching**	• Confirm research topics for pilots • The relevant B/Ds examine the data of the systems they managed that are related to the research topics for pilots • Liaise with relevant B/Ds for data alignment exercise and de-identifying procedures • Check and clarify with the data of the consolidated databases • Set the objectives and outcomes of the analysis • Draft the data analysis methods	11 months
Stage 2 **Data Analysis and Reporting**	• Perform the data analyses • Present the statistical results with charts/ tables; and • Compile a final report of findings.	7 months

[23] The estimated timeline of 18 months for completion of the two pilot projects is premised on the completion of rectification of data inconsistency and alignment of data structure across systems, as set out in Sections 2.1.2.2 and 2.1.2.3 of this document.

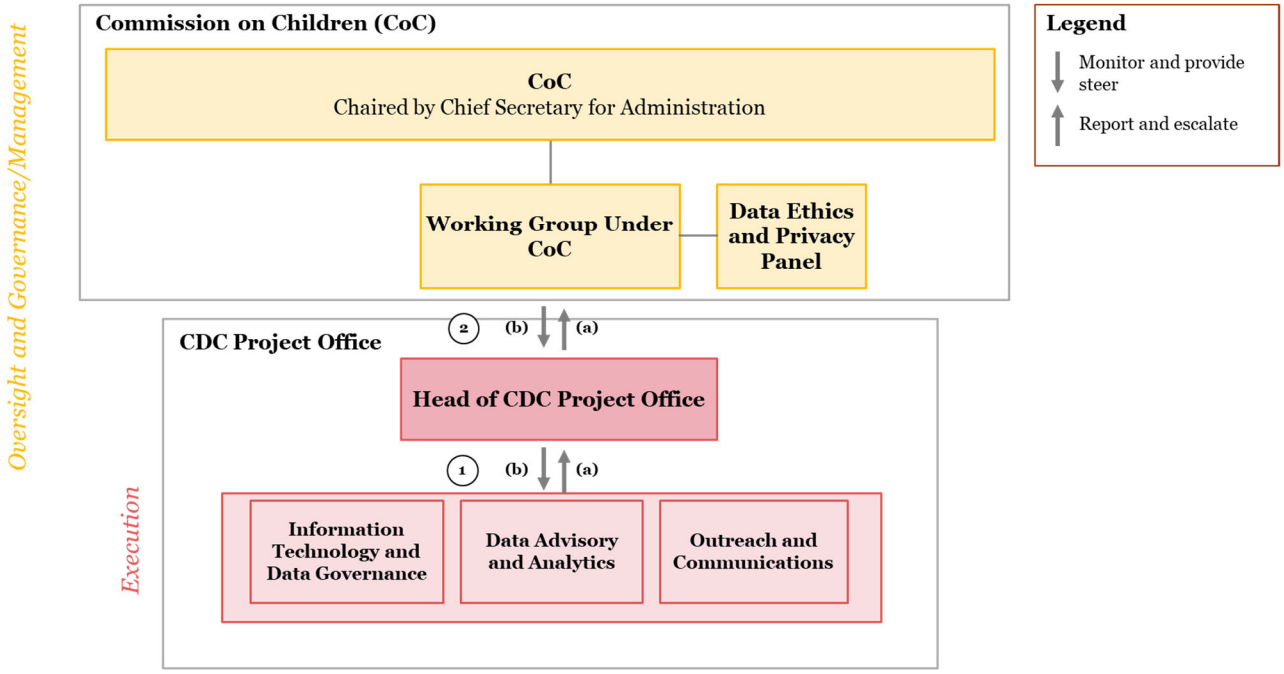
2.1.4.3 Project governance

2.1.4.3.1 Progress Reporting and Quality Review

To ensure effective and timely communication within the various tiers of CDC's governance structure, a clear line of reporting and monitoring shall be set up to allow effective information flow of development updates, project risks and lessons learned.

The indicative reporting and monitoring line for CDC is illustrated below.

Figure 13 Indicative Reporting and Monitoring Line for CDC



A periodic reporting and monitoring mechanism shall also be put in place to ensure clearer allocation of responsibility and allow higher visibility on the work of various parties. An indicative reporting and monitoring mechanism is tabulated below.

Table 34 Indicative Reporting and Monitoring Mechanism for CDC

Item		Objective	Frequency
1	(a)	To report implementation progress and daily operational challenges encountered.	Monthly
	(b)	To manage and oversee the daily operations of the three functions within CDC Project Office, i.e.: • Information Technology and Data Governance; • Data Advisory and Analytics; and • Outreach and Communications.	
2	(a)	To report implementation progress and escalate project risks/ issues to the relevant WG under CoC overseeing CDC if required.	Bi-monthly

Item	Objective	Frequency
(b)	To provide steer on CoC's strategic priorities for children's wellbeing and other Government child-related policies/initiatives;	
	To provide steer on cross-B/D collaboration of CDC such as: • Data alignment and data standards to be adopted by CDC; and • Data sharing to ensure compliance with PDPO and other relevant Government standards; and To endorse and approve future development plans for CDC.	Quarterly

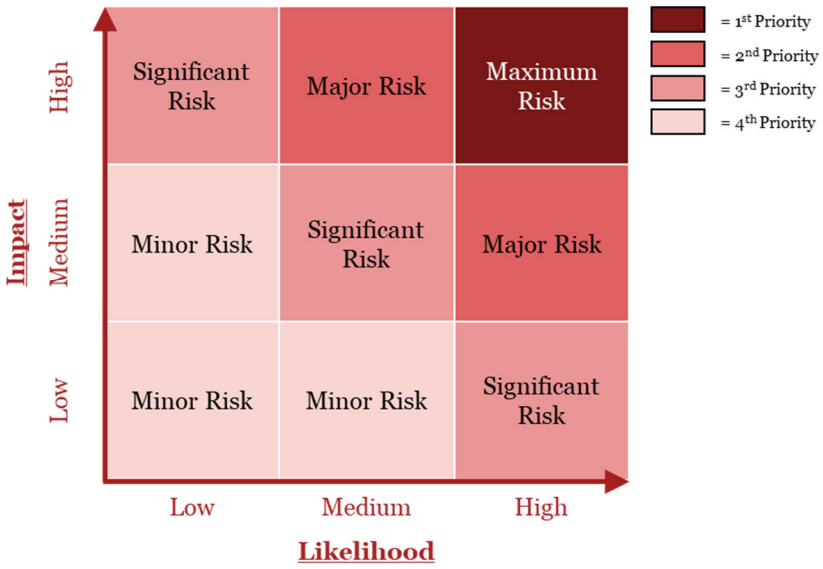
2.1.4.3.2 Risk Management

Effective risk management is a key element to the ultimate success of CDC implementation. Risk assessment is the process of identifying, analysing and prioritising risk events that may occur during implementation. It includes the following elements:

- **Risk Identification**: Risk identification is the process of systematically identifying potential risk events that may affect the implementation of CDC.
- **Risk Analysis**: Risk Analysis is the process whereby the results of risk identification are reviewed. Unlikely risk events are removed from the list, resulting in a more realistic list. These are then analysed to determine both the probability of occurrence and the potential effect on meeting the goals of the programme. Possible risk events are analysed qualitatively and quantitatively. Finally, the risk analysis should determine the best case, worst case and most likely scenario that will happen if the risk event occurs.
- **Risk Prioritisation**: Risk prioritisation is the process of establishing the priority of risk events based on their magnitude of impact and likelihood of happening during the implementation of CDC. The magnitude of impact is defined as follow:
    - **High Impact**: The risk event may impede the overall development of CDC such as potentially attract media attention and lead to public concerns, which shall result in an impact on the image and reputation of CDC and beyond (e.g. image of the Government). For example, security issues (e.g. incidents of data leakage and inappropriate access) and deviation in development from intended purposes and scope of CDC are considered high impact risk events.
    - **Medium Impact**: The risk event may lead to temporary suspension of CDC operations. For example, system outage and major functions unavailable are considered medium impact risk events.
    - **Low Impact**: The risk event may lead to minor disturbances to CDC operations. Some examples of low impact risk events are minor technical fault as well as temporary resource shortage due to expected events such as regular rotations and turnover.

  The risk mitigation prioritisation matrix as illustrated on the next page could be used to determine the priority for mitigation:

Figure 14 Risk Mitigation Prioritisation Matrix



Risk events with high impact and likelihood of happening (i.e. maximum risk) shall be prioritised for risk mitigation before those with medium or low impact and medium or low likelihood of happening.

- **Risk Mitigation**: Risk Mitigation is the process of determining actions or strategies to address risk, reduce or avoid the impact of unwanted events and maximise positive effects on the implementation of CDC. The risk mitigation process involves the development of a risk management strategy and development of the risk management plan. Evaluating and tracking of results of risk management plan execution provides valuable "lessons learnt" information. There are four mitigation strategies:
  - Avoid – plan such that the risk event does not occur
  - Control – plan such that controlled actions are taken in the event of risk event occurrence
  - Retain – plan such that actions are determined when the risk event occurs
  - Deflect – plan such that the consequence of risk event occurrence is borne by a third party

Should risks be realised, issues will need to be resolved on a timely basis. While the division leaders of the CDC Project Office will likely have the authority and/or expertise to resolve issues of a certain nature (e.g. issues of minor risk), other issues will likely require escalation to the WG under CoC overseeing CDC for resolution. A proposed escalation path for implementation issues of CDC, detailing the appropriate level of responsibility for adequate resolution of different types of issues is illustrated in **Table 35** below.

Table 35 Escalation Path for CDC Implementation Issues

Magnitude		Level of responsibility for resolution
Maximum Risk	Business	Working Group under CoC Overseeing CDC
	Technical	
Major Risk	Business	
	Technical	
Significant Risk	Business	CDC Project Office Head
	Technical	
Minor Risk	Business	Division Leaders of CDC Project Office
	Technical	

Eight potential risks for the implementation of CDC were identified based on lessons learned from overseas research and views from stakeholders. A proposed mitigation plan for these potential risks are elaborated in detail on the next page.

Table 36 Potential Risks and Mitigation Strategies of CDC Development

#	Potential risk	Anticipated likelihood and impact			Mitigation strategy(ies)
Business and Legal Dimension					
1.	B/Ds are not able to share data due to regulatory reasons.	Existing regulations (e.g. PDPO and departmental guidelines etc.) may restrict data sharing from certain B/Ds (e.g. HKPF), especially on data pertaining to individual child in child abuse cases; therefore, the likelihood of such a risk is considered to be high. Restricted data sharing from certain B/Ds may hinder certain data linkage projects or longitudinal studies that aim to discover risk factors for child abuse cases; thus, the impact of this risk is considered medium.			• Need for an overall policy and/or legislative support for enabling sharing of data pertaining to children with risk of abuse or neglect • Potential for exemption from PDPO if the identity of child abuse victims are anonymised before sharing to CDC, so as to protect privacy of children with risk of abuse or neglect
		Likelihood	Impact	Priority	
		High	Medium	Major Risk	
2.	Research participants may potentially face unintended consequences, such as labelling effect, caused by data collection targeting a specific cohort of children.	Thematic research, especially longitudinal studies on a specific cohort of children in need, may lead to unintended labelling effect on a certain group of children if the intended research purpose is not clearly conveyed to the data subjects and the public. This could lower public confidence on sharing data and trust in CDC, potentially leading the execution difficulty and suspension of specific CDC projects and impairing the image of CDC. In this regard, the likelihood of such a risk is considered medium, while its impact is considered high.			• Setup of multiple transparent and effective communication channels for CDC to explicitly convey its research purpose and demonstrate the benefits of such studies (e.g. development of associated support programs/policies) to the public to build rapport • Ensure informed consent from data subjects and their parents are obtained and the possible impacts are clearly communicated
		Likelihood	Impact	Priority	
		Medium	High	Major Risk	

<table>
<tr><th>#</th><th>Potential risk</th><th>Anticipated likelihood and impact</th><th>Mitigation strategy(ies)</th></tr>
<tr><td>3.</td><td>CDC Project Office may not be able to secure sufficient resources to operate longitudinal studies should this be considered for execution under the Enhancement Mode.</td><td>Longitudinal studies involve data collection on a specific cohort of children for a well-defined period of time which would require provision of both human and funding resource in a sustained manner. It is anticipated that difficulty to secure sufficient financial and human resources would lead to challenges in executing longitudinal studies. Hence, with a medium likelihood, such risk would pose a medium impact on the operation of CDC.

Likelihood: Medium | Impact: Medium | Priority: Significant Risk</td><td>• Identification of financial and human resource requirements to ensure sufficient funding and human resources before the execution of longitudinal studies
• Policy objectives should be clearly defined to support execution of longitudinal studies
• Clear definition of scope of work (e.g. identification of data to be captured and key stakeholders to be engaged) and relevant research topics for longitudinal studies</td></tr>
<tr><td colspan="4">Data Dimension</td></tr>
<tr><td>4.</td><td>Inappropriate access of data, data leakage or misuse may potentially impact the reputation of CDC and beyond.</td><td>Data security-related risks, such as inappropriate access by unauthorised personnel, accidental data leakage during or before the process of de-identification and linkage, data misuse when sharing data to parties other than the data owners, etc., are considered high-impact risks. Despite a rigid data security system in place to minimise likelihood of these risks, they can potentially lead to failure in addressing data privacy concerns in CDC development and complying with PDPO. Eventually, these risks could severely impede the overall development of CDC and impact the reputation of CDC as well as public trust towards the government.

Likelihood: Low | Impact: High | Priority: Significant Risk</td><td>• Periodic monitoring and assessment of the overall implementation/development of CDC to identify key issues and areas for improvement</td></tr>
</table>

<table>
<tr><th>#</th><th>Potential risk</th><th>Anticipated likelihood and impact</th><th>Mitigation strategy(ies)</th></tr>
<tr><td>5.</td><td>Data alignment/ standardisation may potentially require much longer time and more efforts than expected.</td><td>Identifying variances in data among different data sources and performing standardisation to align the data require much time and significant cross-departmental efforts, depending on the amount of data to be aligned/ standardised and the level of coordination among B/Ds. If data alignment/ standardisation requires much longer time and more financial and human resources than expected, it could hinder the implementation of CDC. Alongside a high likelihood, this risk could pose a medium but long-term impact on the operation of CDC.

Likelihood: High | Impact: Medium | Priority: Major Risk</td><td>• Policy objectives should be clearly defined to drive cross-departmental collaboration
• Robust communication and engagement with all participating stakeholders/data providers (e.g. B/Ds, researchers/academia, practitioners, service providers, etc.) to ensure the clarity of data specifications, appropriate data quality, and thorough understanding of data source, purpose, and context to reach consensus on data alignment</td></tr>
<tr><td colspan="4">Technical Dimension</td></tr>
<tr><td>6.</td><td>B/Ds may have different technical readiness and capability in performing data sharing with CDC.</td><td>Technical readiness and capability for data sharing (e.g. data captured in electronic or in paper forms, quality of data, etc.) may vary across the databases of different B/Ds. While the early stage of CDC development could be kick-started with databases with higher technical readiness and capability for data sharing, those with lower technical readiness and capability may delay future development/ enhancement of CDC if the technical readiness is not sufficient for data sharing with CDC. Hence, the likelihood for this risk is considered high with a medium impact.

Likelihood: High | Impact: Medium | Priority: Major Risk</td><td>• CDC Project office to provide technical support to B/Ds with a lower level of technical readiness and capability to ensure compatibility for data sharing
• Regular checkpoint on participating B/Ds databases to ensure alignment of technical readiness and capability</td></tr>
</table>

<table>
<tr><th>#</th><th>Potential risk</th><th>Anticipated likelihood and impact</th><th>Mitigation strategy(ies)</th></tr>
<tr><td colspan="4">Implementation Dimension</td></tr>
<tr><td>7.</td><td>The public may potentially misunderstand the intended purposes of CDC.</td><td>Lack of communication with the public may lead to misalignment on the understanding of the intended purposes of CDC (e.g. data collection/ sharing for the purpose of facilitating coordination among B/Ds on service provision and enhancing efficiency for internal processing), impacting public trust towards its establishment and work. Despite a medium likelihood, the impact of such a risk is considered high, as it could impact the overall implementation and reputation of CDC.
<table><tr><th>Likelihood</th><th>Impact</th><th>Priority</th></tr><tr><td>Medium</td><td>High</td><td>Major Risk</td></tr></table></td><td>• Regular public engagement for the purpose of collecting feedback, understanding expectation and providing periodic updates on the progress of CDC to ensure alignment with the public on the intended purpose of CDC
• Publicity promotion to demonstrate the objective and benefits of CDC to the public to align expectation and enhance public confidence</td></tr>
<tr><td>8.</td><td>Users may lack the knowledge/ skills required for using CDC, leading to a low adoption rate.</td><td>Without proper training or comprehensive instructions provided to data users, these users could be discouraged from using CDC if they do not know how to use the system at the first place, potentially leading to a low adoption rate of CDC. This could impact the continued development of CDC. Therefore, the likelihood of such a risk is considered medium, while its impact is considered high.
<table><tr><th>Likelihood</th><th>Impact</th><th>Priority</th></tr><tr><td>Medium</td><td>High</td><td>Major Risk</td></tr></table></td><td>• Involvement of all users when designing CDC and during the requirements gathering process; and identification of key users for User Acceptance Test (UAT)
• Provision of user training after the launch of CDC to make sure users are capable of using CDC
• Release of a set of thorough and informative users' guideline to all data users
• Regular review and update of the users' guideline</td></tr>
</table>

2.1.4.3.3 Stakeholder Engagement and Change Management

Stakeholder Engagement

To ensure the CDC development is fit-for-purpose and to build public trust in CDC, the importance of having targeted and on-going stakeholder engagement was emphasised in our findings from overseas research and stakeholder consultation. The objective of adopting a comprehensive stakeholder engagement plan is to allow appropriate information to be disseminated to various stakeholder groups and strengthen transparency and visibility of CDC development across these stakeholder groups.

As identified in **Section 4.2.1.4** of the Third Progress Report, key stakeholder groups to consider for engagement include children's workforce (e.g., social workers, educators and healthcare practitioners etc.), NGO service providers, academics and experts, parents/ guardians, children, and the general public. Discussion topics in stakeholder engagement should vary based on the roles and characteristics of stakeholder groups. Different formats for engagement could also be considered depending on the size, diversity and level of impact of each stakeholder group.

As tabulated on the next page, an indicative stakeholder engagement plan was formulated with consideration of the roles and characteristics of the key stakeholder segments. The objectives, key discussion topics and the corresponding channels for engagement were also detailed in the below table.

In addition to the above stakeholder engagement plan, independent public figures could also be considered as spokespersons or ambassadors of CDC to further improve public perception and increase publicity of CDC development. By leveraging the influence of the spokesperson alongside active public engagement, CDC's benefits, values and objectives could be promoted to enhance public acceptability and trust in CDC.

Table 37 Indicative Stakeholder Engagement Plan

	Potential external CDC users	
	Children's workforce, NGO service providers, academia and experts, and key stakeholders from the general public (e.g. parents/ guardians and children)	**Wider community**
Pre-implementation		
Objectives	• To brief purpose and upcoming involvement required for the implementation of CDC	• To brief purpose and potential benefits for the implementation of CDC
Key discussion topics	• Background of CDC development • Role of stakeholder group and upcoming support required	• Background of CDC development • Potential benefits for CDC
Mode(s) of engagement	• Briefing sessions	• Online and offline advertisement (e.g. physical leaflets, official videos and official publications on websites etc.)
Frequency	• One-off	• One-off
During implementation		
Objectives	• To gather business requirements for the implementation of CDC • To understand technical readiness of stakeholder group in data sharing • To identify potential challenges on data sharing	• To provide regular updates on CDC development
Key discussion topics	• Functionalities required in CDC design • Interfacing/ data sharing requirements • Current technical readiness and required support • Potential challenges in data alignment, data processing and data privacy considerations	• Status of CDC development • Anticipated next steps of CDC development

Mode(s) of engagement	• Workshops • One to one discussion session, if required	• Official publications on Government websites
Frequency	• One-off (multiple rounds during SA&D phase of CDC implementation) • One to one discussion session, if and when required	• Quarterly/ bi-annual
Post-implementation		
Objectives	• To solicit feedback upon completion of each development milestone of CDC	• To solicit feedback from the wider community on the implementation of CDC and the pilot projects
Key discussion topics	• User feedback on CDC usage and adoption • Feedback on demonstrated benefits of pilot projects • Suggestions on topics for further research projects	• Feedback on pilot projects and introduction of CDC
Mode(s) of engagement	• Workshops • One to one discussion session, if required • Online communication channels	• Online communication channels • Community-led engagement
Frequency	• One-off for workshops (multiple rounds after each development milestone of CDC) • One to one discussion session, if and when required • On-going for online communication channels	• On-going

Change Management

While the stakeholder engagement plan concerns communication with possible external users of CDC, it is also important to have a structured process to prepare and support internal CDC users in relation to the implementation of CDC. Internal users, including data-contributing B/Ds, should be engaged throughout the various stages of CDC implementation to allow for two-way communication: to foster B/Ds' understanding and involvement, as well as to collect their feedback for facilitating the CDC development and their adjustment.

An indicative change management plan is tabulated below, setting out the respective objectives, key discussion topics and the corresponding communication channels. In particular, the discussion topics covered corresponds to the anticipated involvement of B/Ds during different stages of CDC development.

Table 38 Indicative Change Management Plan

Pre-implementation	
Objectives	• To brief purpose and upcoming involvement required for the implementation of CDC
Key discussion topics	• Background of CDC development • Role of B/Ds and upcoming support required
Mode(s) of communication	• Briefing sessions
Frequency	• One-off
During implementation	
Objectives	• To gather business requirements for the implementation of CDC • To understand technical readiness of B/Ds in data sharing • To identify potential challenges in data sharing
Key discussion topics	• Functionalities required in CDC design • Interfacing/ data sharing requirements • Current technical readiness and required support • Potential challenges in data alignment, data processing and data privacy considerations
Mode(s) of communication	• Workshops
Frequency	• One-off (multiple rounds during SA&D phase of CDC implementation)

Post-implementation	
Objectives	• To solicit feedback upon completion of each development milestone of CDC
Key discussion topics	• User feedback on CDC usage and adoption • Feedback on demonstrated benefits of pilot projects
Mode(s) of communication	• Workshops • Online communication channels
Frequency	• One-off for workshops (multiple rounds after each development milestone of CDC) • On-going for online communication channels

2.1.4.3.4 Project Constraints

Project constraints are limitations that would need to be taken into account so as to ensure that the project is executed with quality, in accordance with the schedule, and without causing budget overrun. Based on the proposed PMP, the project constraints concerning the development of CDC are categorised and elaborated in the table below.

Table 39 Project Constraints of CDC

Areas	Details
Time Constraint	Adopting a phased approach for implementation, certain key items of CDC development have to be completed within a specified timeframe. For instance, a full data ethics and privacy panel needs to be established before the transitioning to the Enhancement Mode while the pilot studies will provide recommendations on data to be prioritised for alignment. This dependency suggests that the delay of one individual task could affect the schedule of the next task, as well as the execution of later phases of CDC. To ensure the timely completion of key tasks, regular reporting should be performed within the CDC Project Office and to the WG under CoC overseeing the CDC.
Scope Constraint	Projects spanning across a long period of time are more susceptible to issues such as features or functions that are no longer relevant and/or having enhancement that might deviate from its key objective development. Changes in scope could have unintended implication on the project's schedule and cost. Given that CDC is a long-term project, the scope for CDC development needs to be clearly defined at the onset and communicated in a timely manner to B/Ds involved. The CDC Project Office should also maintain close liaison with B/Ds to ensure the development of CDC is within the defined scope. Nonetheless, when developing the CDC, flexibility should be exercised to take into consideration new IT development or trends in children issues, both of which may entail the need for changes to the agreed scope.
Quality Constraint	Considering the involvement of multiple parties in the development of CDC, such as the participation of multiple data-contributing B/Ds, it is possible that different parties may have differing ability in meeting the

Areas	Details
	requirements/needs for collaborative projects (e.g. for macro-trend analysis and/or data linkage purposes). Having regard to the number of parties involved in CDC development, there is a need to define the key milestones to keep all parties on the same page and work toward the defined scope. In addition, a monitoring mechanism may also be put in place between the CDC Project Office and the B/Ds to ensure the quality of input from all parties involved.
Resources Constraint	The availability of required resources is crucial to CDC development. Of particular importance is the availability of personnel for carrying out the functions of the CDC Project Office. In staffing the CDC Project Office, careful consideration should be given to whether the selected personnel possess relevant expertise and working experience. Moreover, it is envisaged that the IT implementation and the pilot studies will be commissioned to third parties outside of the Government. Dependency on external parties for delivery of services may impact the overall schedule of CDC development. Appropriate personnel should be selected to perform project management on these commissioned services in order to ensure quality delivery and adherence of third parties to project schedule.
Risk Constraint	Risk refers to unexpected occurrences which could impact project implementation. In the context of CDC, some significant risks may include inflation risks given the duration of CDC implementation and data security-related risk which may affect public acceptance and buy-ins of CDC. Risk management strategies should be adopted in responding to risks. On this, a risk management framework is detailed at **Section 2.1.4.3.2**.

2.1.4.3.5 Critical Success Factors

Critical Success Factors (CSFs) refers to areas that are crucial and necessary for a project's achievement of its mission and purpose. Based on the proposed PMP, the CSFs for CDC development are identified below.

Table 40 Critical Success Factors of CDC

Areas	Critical Success Factors (CSFs)
Governance	• An established mechanism for reporting and review; and • Timely risk escalation within the CDC Project Office as well as between the CDC Project Office and the WG under CoC overseeing CDC.
Expertise and Capabilities	• Expertise of members of the oversight bodies in relevant subject areas (e.g. children's well-being, data analytics, data standards); • Skills and expertise of the staff of the CDC Project Office (e.g. data governance, research, statistical analyses); and

Areas	Critical Success Factors (CSFs)
	• Capability to create a de-identification key for facilitating de-identification of data for pilot and linkage projects.
Internal Collaboration	• Effective communications on involvement required with data-contributing B/Ds; and • Close and constant liaison with and provision of support to B/Ds in relation to CDC development.
External Buy-in	• Effective communication with the public on the purposes and development progress of CDC; • Satisfactory completion of Privacy Impact Assessment to instil public trust; and • Successful implementation of and insights from pilot projects to demonstrate the benefits of CDC.
Privacy and Ethics	• Compliance with PDPO and effective execution of safeguards for privacy protection and ethical data usage.

2.2. Qualitative Benefits for CDC's Implementation

Potential Benefits Arising from the Development of CDC in Hong Kong

With reference to stakeholder engagement findings, overseas experience and further desktop research, it is anticipated that the development of the initial CDC model has the potential to bring about the following direct benefits to key users.

Table 41 Potential Direct Benefits of CDC for Key Users

Potential Direct Benefits	CDC Model			Key Users			
	Foundat-ional Mode	Enhancement Mode		Governm-ent	Academia	Children's Workforce	General Public
		Data Linkage	Longitudin-al Studies				
Provide an integrated and comprehensive view on current state of children's well-being	√			√	√	√	√
Raise awareness of emerging trends on child-related issues	√			√	√	√	√
Enhance efficiency in the search for credible and reliable aggregate data concerning children	√				√	√	√
Enable better design of measures for tracking children well-being in the future	√			√			
Facilitate policy review and service planning	√	√	√	√		√	
Enable the identification of risk factors for child-related issues		√	√	√	√	√	
Facilitate multistakeholder collaboration	√	√	√	√	√	√	

It is to note that the extent to which these direct benefits will be realised depend on the development objective of CDC development. For example, the development objective "causal analysis and risk identification" under the Enhancement Mode is likely better positioned in realising the benefit of supporting policy review and children's service planning than "trend-monitoring at a macro level" under the Foundational Mode, given the greater depth of insights that could be generated by performing data linkage and longitudinal studies.

Further elaboration of the benefits is detailed below:

- **Provide an integrated and comprehensive view on current state of children's well-being –** Provision of aggregate data under the Foundational Mode of CDC could provide a strong foundation for generating a holistic view on the state of children's well-being in Hong Kong. This may either be on children in general or on key priority areas. In the case of Children's Headline Indictors (CHI) in Australia, its data (e.g. overweight and obesity, child abuse and neglect, social and emotional well-being, etc.) was used by the Department of Health in their comprehensive review of the state of children's health and well-being in Australia during the formulation of the National Action Plan for the Health of Children and Young People 2020-2030.

  Furthermore, the use of a single source of data could facilitate benchmarking, with a view to understanding and identifying areas of variability, especially in relation to the areas of "health and safety" and "education and skills". Aggregate statistics available in CDC could serve as territory-wide baselines, of which reference could be made in benchmarking within Hong Kong (e.g. between districts, segments of children). For example, by making comparison to the national norms (i.e. consolidation of national statistics published by the Canadian Government), the Early Child Development Mapping Project (ECMap) initiative was able to identify developmental areas that Albertan children experienced great difficulty in. In addition, the territory-wide statistics could also serve as a point of reference against other overseas jurisdictions. For instance, by leveraging national norms of various countries, UNICEF Canada found a higher rate of obesity amongst young people in 2014.

- **Raise awareness of emerging trends on child-related issues –** Reporting of aggregate data for monitoring macro-trends under the Foundational Mode has the potential to highlight existing and/or emerging issues that require further investigation. For example, data (e.g. life satisfaction and happiness yesterday) from the UK Children's Wellbeing Measures was referenced in a report from the UK's Department of Education to raise awareness on the changing trends in the well-being of children and young people, to prompt questions and identify further research pathways for investigation as well as to support intervention.

  The demonstration of territory-wide children's statistics with the availability of further data breakdown could also bring awareness on challenges experienced by specific segments of children. In the case of Australia Well-being Monitoring Framework (WMF), the finding that Aboriginal young offenders were 20 times more likely to be in detention than non-Aboriginal young people in 2012 highlights a differential treatment of Aboriginal youth in the criminal justice system.

- **Enhance efficiency in the search for credible and reliable aggregate data –** Through the provision of a data catalogue for enhanced data discovery under the Foundational Mode, there is potential for CDC development to reduce the search time for credible and reliable children's data in areas related to "health and safety" and "education and skills". It was observed in the European Union (EU) that the benefits of open data include time savings for users of applications using open data[24]. In particular, open data could improve the speed at which research can be conducted and disseminated.
- **Enable better design of measures for tracking children well-being in the future –** CDC's function of macro-level trend monitoring could allow Government B/Ds to have a better understanding of the trends concerning issues related to children and a more comprehensive view of children well-being whereby different well-being domains that fall under the purview of different B/Ds are duly considered. Enhanced understanding of the current landscape and changing needs of children has the potential to facilitate B/Ds in developing more appropriate measures for tracking children well-being in the future. These measures could enable systematic monitoring of individual programme performances as well as the Government's overall progress towards meeting set goals for enhancing children well-being. For example, Australia's CHI provided data and statistics to support tracking of Queensland's Child Protection System. Selected indicators from the CHI were used to keep track of Queensland's performance and progress towards improving outcomes as well as meeting the national goals relating to child protection system. In addition, disaggregation of data by indigenous status allows an emphasis placed on Indigenous children and families, illustrating any discrepancies in well-being indicators.
- **Facilitate policy review and service planning –** the Government and Children's workforce alike could benefit from better application of data, enabled by the use of CDC, in reviewing their policies and planning services.
    - **Foundational Mode:** Reporting of macro-trends with breakdown could enable better monitoring of trend that are of high strategic priorities, facilitate understanding on the plight of specific segments of children and serve as reference material/evidence base for policy action where relevant. For example, the statistic on the time spent by children on social network in the UK Children's Well-being Measures sheds light on the changing trends on children's social media usage. This finding was used as an evidence base to support the UK government's review of existing regulatory landscape for safeguarding the well-being of children online and to bring forward an online safety legislation that covers the full range of online harms. Furthermore, in the case of Australia's WMF, disaggregation of data by

[24] https://data.europa.eu/sites/default/files/edp_creating_value_through_open_data_0.pdf https://data.europa.eu/sites/default/files/the-economic-impact-of-open-data.pdf

Aboriginal status enabled local authorities a glimpse into the specific challenges faced by Aboriginal youth, calling for a policy review. In particular, the statistics of WMF indicated an overrepresentation of Aboriginal youth in detention in the criminal justice system when compared with non-Aboriginal youth. This, along with other reference materials, prompted the Department of Correction Services to review the degree of disadvantage faced by the Aboriginal people and consider legislative means of reducing the rate of overrepresentation of Aboriginal youth in the criminal justice system. Subsequently, a new Order to create a non-government managed, culturally appropriate community approach towards Aboriginal youth referencing WMF statistics was proposed.

- **Enhancement Mode:** Data linkage projects and longitudinal studies allow for more robust analyses, of which findings could be used by key users to inform resource allocation during service planning. For example:

  - ***Data linkage projects*:** Data linkage across otherwise siloed services can further support service design and delivery, by using population-level evidence of the need for supportive, intensive and targeted services. For example, by linking data from the South Australian Early Childhood Development Project (ECDP) (i.e. data in relation to perinatal, births registration, housing and child protection) and Child and Family Health Services (CaFHS), CaFHS were able to understand different levels of adversity and vulnerability (i.e. measured the number of socioeconomic, trauma, psychosocial and health-related risk factors) experienced by infants across South Australia. The disaggregation of data available also highlighted the higher levels of vulnerability experienced by Aboriginal and Torres Strait Islander communities due to the historical forces creating multiple forms of discrimination and adversity. This research directly informed the development of CaFHS' new model of care and, through small-area-based adversity profiling, is also informing resource allocation from lower to higher areas of need based on the absolute number of expected births at different adversity levels.

  - ***Longitudinal studies*:** In the longer term, should longitudinal studies be considered under the Enhancement Mode, it could also facilitate better policy review and service planning through tracking the same cohort of children over time. For instance, in Canada, a more granular set of data, made available through a five-year longitudinal study tracking a cohort of kindergarten children, allowed for a more robust analysis on the development differences amongst children in different communities of Alberta. This led to the identification of incidences whereby children living in the rural Strathcona County were experiencing greater difficulty in one or more of the developmental areas. These targeted findings directly informed the decision-making of some government bodies on facilities

improvement, such as local parks and recreation departments in resource planning. For example, Strathcona County established a free play-based space between 2012 and 2013, which subsequently led to a major investment in the Love to Play Room in the Country.

- **Enable the identification of risk factors for child-related issues –** In enabling robust analyses, data linkage projects and longitudinal studies under the Enhancement Mode of CDC could support the identification of risk factors with high level of validity and applicability for policy making. Risk factors identified for child-related issues could help researchers identify further research pathways for investigation; support the Government in their policy and investment decisions; and support children's workforce in their identification of children requiring additional support as well as early detection and intervention.

    - ***Data linkage projects*:** By linking child protection data from ECDP and data from the Australian Early Development Census, it was found that children at age 5 who experienced higher levels of contact with the child protection system were more likely to experience developmental vulnerability on 1 or more domains (i.e. physical health and well-being, social competence, emotional maturity, language and cognitive skills and communication skills and general knowledge). Particularly noteworthy was the finding that children who have ever been notified (not screened in), and never had any more serious child protection contact, are nearly twice as likely to be developmentally vulnerable at age 5. This research suggests that any level of contact with child protection, even notifications determined not to be 'child protection matters', is an indicator of high risk for poor child development. The finding disputes the anecdotal evidence that many initial notifications are 'not real' and echoed the need for a whole-of-government coordinated response to address children protection-related concerns.

    - ***Longitudinal studies*:** Through tracking the development of a cohort of children in Alberta over the course of 5 years, the ECMap initiative found low socio-economic levels were associated with poor early development outcomes. Where positive socio-economic levels were present, the proportion of children doing well is greater. These findings suggest that low or less favourable socio-economic status increases the risk for experiencing poorer early childhood development outcomes.

- **Facilitate multi-stakeholder collaboration –** The set-up of a CDC offers collaborative opportunities between the Government, academia, and NGOs to increase the knowledge pool in developing tools, intervention and programmes to promote children's well-being. Academia and NGOs use more quality data from the Government in supporting their research and services, whereas the Government could leverage additional insights and expertise from the sector in formulating applicable policies. For example:

    - **Foundational Mode:** The findings of the Children's Well-being Measures published by the Office for National Statistics (ONS) of the UK government

provide reference points for NGOs working on children matters in advocating the use of subjective indicators in measuring well-being of children. Selected measures formed reference materials for Coram Voice, a children's rights charity, and the School of Policy Studies of the University of Bristol to build upon their own research in developing a subjective measure (i.e. survey) for looked after children (i.e. children having been in the care of local authority for more than 24 hours). The measure was piloted with the support of six local authorities, leading to further refinement and validation.

- **Enhancement Mode:** In undertaking ECMap, Alberta Education of Canada collaborated with experienced academia, such as the Offord Centre for Child Studies at the MacMaster University in Hamilton (i.e. the creator of the Early Development Instrument (EDI)), to adopt the EDI as the key measurement tool for conducting research and preparing reports. Alberta Education also contracted work to Community-University Partnership for the Study of Children, Youth and Families (CUP), based at the University of Alberta, to leverage their expertise in research (e.g. data analysis) and community engagement. In addition, community coalitions in different zones formed by members from diverse background (e.g. healthcare, education, parents and other community members) were funded facilitate the collection of primary research data for ECMap, define the geographically boundaries to ease operation of the project as well as to disseminate research output in the community.

Glossary

Application Programming Interface (API)

Attention-deficit/hyperactivity disorder (ADHD)

Autism Spectrum Disorder (ASD)

Constituency areas (CA)

Cancer Statistics Query System – Children and Adolescents (CanSQS).

Case Management and Investigation System (CMIS)

Census and Statistics Department (C&SD)

Central Databank on Children (CDC)

Central Referral System for Residential Child Care Services (CRSRC)

Child and Family Health Services (CaFHS)

Child Assessment Service Information System (CASIS)

Children's Headline Indicators (CHI)

Child Health Service System (CHSS)

Child Protection Registry (CPR)

Clinical Data Analysis and Reporting System (CDARS)

Commission of Children (CoC)

Community-University Partnership for the Study of Children, Youth and Families (CUP)

Data Quality Assessment Framework (DQAF)

Department of Health (DH)

Department of Justice (DOJ)

District Council (DC)

Education Bureau (EDB)

Early Childhood Development Mapping Project (ECMap)

Early Childhood Development Project (ECDP)

Early Development Instrument (EDI)

Electronic identity (eID)

Emotional and Behavioural Difficulties (EBD)

Generic Statistical Business Process Model (GSBPM)

Government Backbone Network (GNET)

Government Cloud Infrastructure Services (GCIS)

Government Records Service (GRS)

Hearing impairment (HI)

Hong Kong Jockey Club (HKJC)

Hong Kong Police Force (HKPF)

Hospital Authority (HA)

Hospital Authority Data Collaboration Laboratory (HADCL)

Information and Technology (IT)

Infrastructure-as-a Service (IaaS)

Intellectual disability (ID)

Kindergarten Education Scheme System (KGESS)

Labour and Welfare Bureau (LWB)

Multi-disciplinary Case Conference (MDCC)

Mental illness (MI)

Motor impairment (MotorI)

National Statistician's Data Ethics Advisory Committee (NSDEAC)

Non-Governmental organisations (NGOs)

Office of the Government Chief Information Officer (OGCIO)

Office of the Privacy Commissioner for Personal Data (PCPD)

Personal Data (Privacy) Ordinance (PDPO)

Personal Information Collection Statement (PICS)

Physical disability (PD)

Project Management office (PMO)

Project Management Plan (PMP)

Public Sector Information (PSI)

Virtual machines (VM)

System analysis and design (SA&D)

Social Welfare Department (SWD)

Special Education Management Information System (SEMIS)

Special Educational Needs (SEN)

Specific Learning Difficulties (SpLD)

Speech impairment (SI)

Student Information Management System (STIMS)

System for Managing the Assessment of Student Health (SMASH)

The Data Documentation Initiative (DDI)

the International Classification of Disease, Ninth Revision, Clinical Modification (ICD-9-CM)

The University of Hong Kong (HKU)

User Acceptance Test (UAT)

Visceral disability and chronic illness (VD&CI)

Visual impairment (VI)

Well-being Monitoring Framework (WMF)

Working Group (WG)

Appendix A - Summary of Views Gathered from Stakeholder Engagement

This appendix compiled key findings on the views and expectations of key stakeholders in relation to the feasibility of developing a CDC. Views are gathered from the engagement activities conducted in Phase 3 of the Study including interviews, focus group discussions (FG), engagement sessions (ES) and a survey targeting at members of the general public.

Interviews

Organisations	Completed
Government Bureaux/Departments (e.g. Social Welfare Department, Education Bureau, Department of Health, Hospital Authority, etc.)	12
NGOs (e.g. Hong Kong Jockey Club, Hong Kong Society for the Protection of Children, etc.)	4
Academia/Individuals (e.g. School of Public Health/HKU, Faculty of Law/CUHK, etc.)	4
Total	**20**

Focus Groups (16 completed with 64 organisations attended)

Categories	Attended
Social Welfare, Family Groups, Schools/Educators (e.g. Hong Kong Christian Services, Heep Hong Parents' Association, etc.)	49
Social Science	8
Healthcare	7
Total	**64**

Engagement Sessions (5 completed with 112 people attended)

Participants	Attended
ES#1 Children (aged 13-17)	21
ES#2 Ethnic Minorities (aged13-17)	18
ES#3 Parents of Children (aged at or below 12)	32*
ES#4 Parents of Children with SEN aged below 18	21
ES#5 General Public (aged above 18)	20
Total:	**112**

*A total of 11 children aged between 6 to 12 participated ES#3 together with their parents.

Survey

There are 1 007 respondents for the survey.

Key highlights of stakeholders' views on the areas of Development Objectives, Scope of Data, Public Trust Building and Data Ethics and Implementation of CDC are tabulated below –

Key Highlights of Stakeholders' Views
Development Objectives
• "Monitoring children's well-being at a macro-level" should be a development objective of CDC • There is a need to identify more specific topics or provide breakdown of aggregate data for more meaningful macro-level analysis • Many stakeholders agreed that "supporting causal analysis and risk identification" should be considered as a development objective. • Divergent views on the approach (i.e. Option A – Collecting data for longitudinal studies and Option B – Linking existing administrative data of children) in meeting "causal analysis and risk identification" . • Whilst stakeholders could see the value in "case tracking to enhance coordination of the children's workforce", they generally agreed that this objective would trigger serious privacy concerns. • Stakeholders see the potential to track specific segments of children under limited use and with consent. • Stakeholders generally expect CDC development to bring about benefits for different key parties, i.e. the Government, academia and children's workforce, and general members of the public
Priority Areas
• Many stakeholders were of the view that Children with Risk of Abuse or Neglect and Children with SEN should be considered as priority areas • Key stakeholders' views in relation to the two priority areas are broadly summarised below: <u>Children with Risk of Abuse or Neglect</u> o This segment is commonly highlighted by stakeholders engaged through interviews and ES. o Over 70% of respondents (71.3%) selected this segment as a priority area in the Survey. o Stakeholders engaged through different modes expressed concerns in view of the increasing occurrence of child abuse and fatal cases over the past years. A few

stakeholders also opined that this trend has exacerbated amidst the ongoing COVID-19 situation.

- Many stakeholders were of the view that there would be a need to generate deeper insights in order to identify susceptible cases and provide target support.

Children with SEN

- This segment is commonly highlighted by stakeholders engaged through interviews, FG and ES.
- More than half of respondents (56.3%) selected this segment as a priority area in the Survey.
- Stakeholders engaged through different modes were of the view that there would be a need to focus on this segment due to the lack of data available in understanding the current situation. Current provision of services to SEN children is also perceived to be inadequate.
- There were views that the ongoing COVID-19 situation could have a negative impact on the well-being of SEN children.
- A few stakeholders engaged through different modes suggested prioritising certain types of SEN conditions such as Autism Spectrum Disorder (ASD) and Attention Deficit Hyperactivity Disorder (ADHD). However, a stakeholder from FG also cautioned that prioritising specific SEN types may lead to public perception of inequitable treatment across SEN types.

Scope of Data

On data domain

- Stakeholders generally agreed that there would be a need to focus on more than one data domain in understanding children's well-being, with "health and safety" and "education and skills" most mentioned

On data type and data sharing arrangement

- Stakeholders generally agreed that aggregate data could be shared through CDC, with sharing of essential identifiable data triggering the most privacy concerns. There were more divergent views in the sharing of other forms of data

On data governance

- Stakeholders generally agreed on the need for multi-level safeguards for protecting the privacy and security of data

On CDC architecture and design

- A decentralised architecture for CDC development is perceived to be least disruptive to database owners
- Many stakeholders envisioned CDC to be a platform where children-related data, information and analysis would be available
- Several stakeholders made suggestion on the features and technology to be adopted by CDC
- Stakeholders generally agreed that existing official data from relevant B/Ds could serve as key sources of data for CDC, although other sources of data (i.e. collected through both primary and secondary means) should also be considered
- Stakeholders generally agreed on the need to set access rights for different stakeholder groups

Public Trust Building and Data Ethics

On public trust building

- Many stakeholders highlighted data privacy and security as key public concerns in CDC development
- Stakeholders generally agreed that public engagement is a pre-requisite for instilling public trust in CDC implementation, and recognised the below as key stakeholders –
    - Children workforce (e.g. social workers, educators and healthcare practitioners)
    - NGO service providers
    - Academics and experts:
    - Parents/guardians:
    - Children:
    - Public:
- Several stakeholders suggested on ways to improve public perception of CDC development including engaging public figure(s) with positive image and engaging public relations specialists for public communications

On data ethics

- When compared with concerns over data privacy and security, very few stakeholders emphasised data ethics as a prioritised concern

- Key suggestions on considerations for data ethics are broadly summarised below:
    - Formulating a clear data ethics framework governing data usage (e.g. setting principles and guidelines for use);
    - Communicating the data ethics framework to the public to gain public trust;
    - Establishing a data ethics committee (comprising independent parties with no conflict of interest and with relevant subject matter knowledge);
    - Establishing a set of protocols for assessing study designs, method of collecting data and usage of data;
    - Consulting the Privacy Commissioner for Personal Data for specific advice pertaining to ethical and privacy issues; and
    - Establishing safeguards such as the need for pre-requisite consent, setting access controls and adopting an audit mechanism (refer to "Data Governance" above).

Implementation Approach

- Stakeholders generally agreed that a phased approach for CDC implementation should be adopted although considerations would need to be given on a publicly acceptable timeline
- Key suggestions regarding implementation priority are broadly summarised below:
    - Need to define objectives, priority areas and assess privacy implications before undertaking other tasks of designing and implementing CDC;
    - Potential to conduct trend analysis derived from aggregate and anonymised data as a foundation before embarking on research of specific topics;
    - Need to implement a pilot phase with development of use cases to demonstrate benefits of CDC with prioritisation of specific segments of children such as Children with Risk of Abuse or Neglect and Children with SEN;
    - Consider aggregating more readily available data first (i.e. suggestion that health and education data would be more ready) before release of other data that is more sensitive (e.g. further breakdowns or identifiable data) or require consent for further collection;
    - Consider classifying existing data into categories to better organise data for discovery;
    - Consider consolidation or linkage of Government-owned data first before extending to other sectors;
    - Consider involvement of key B/Ds that hold the most relevant children-related data within the Government first; and

  - Consider setting up an online platform for public communications during the pilot phase.

- Several stakeholders provided suggestions on pilots to be considered for demonstrating the benefits of CDC, such as:

  - Release of thematic reports on children in general or specific segments (e.g. profile of child mortality in Hong Kong, current state of children with SEN, impact of COVID-19 and social distancing measures on mental and physical wellness of children in general or on specific segments of children); and

  - Release territory-wide child development data (e.g. aggregate data on child development in early years) for researcher to benchmark/reference with overseas jurisdictions.

Appendix B – Summary of Findings from Review of Overseas Practices

Project	United Kingdom – Children's Well-being Measures	United Kingdom – ContactPoint	Australia – Children's Headline Indicators (CHI)	Western Australia – Well-being Monitoring Framework (WMF)	Alberta, Canada – Early Child Development Mapping Project (ECMap)
Key Role	Nation-wide measurement for better reporting on quality of life based on children's data. Its aim is to establish measures of national well-being that adequately reflect the needs of children. Its target segment is children aged 15 and below.	An online directory for connecting national and local sources of children's contact data. Its aim is to enhance coordination of the children workforce by sharing children's information for case tracking purposes. Its target segment is general children up to the time they reached their 18th birthday.	A nation-wide measurement for guiding strategy and policy. Its aim is to help guide and evaluate policy development by measuring progress on a set of indicators. Its target segment is children aged 12 and below.	A state-level measurement framework for guiding strategy and policy comprising three components. Its aim is to provide information on a range of indicators on the well-being of children and young people across the life course overtime. Its target segment is children aged below 18 years old.	A five-year longitudinal study focusing on early childhood. Its aim is to analyse the Early Development Instrument (EDI) in its main aspects by attempting to portray developmental outcomes of Alberta's kindergarten using national benchmarks. Its target segment is children aged 5 and below.

Project	United Kingdom – Children's Well-being Measures	United Kingdom – ContactPoint	Australia – Children's Headline Indicators (CHI)	Western Australia – Well-being Monitoring Framework (WMF)	Alberta, Canada – Early Child Development Mapping Project (ECMap)
Development Timeframe	11 years, since conceptualisation phase in 2009	N.A. (Project was terminated upon initial implementation stage)	15 years, since conceptualisation phase in 2005	13 years, since 2007	11 years, since launch in 2009 (including 5 years for data collection followed by ongoing community engagement)
Governance	Oversight by the board of a public body focussing on statistical reporting, with operations undertaken by an executive arm.	Oversight by a government agency with project management undertaken by an operating unit within the agency.	Oversight by a high-level strategic committee, with operations undertaken by a public body with a focus on health and welfare statistics.	Oversight by a joint standing committee, with operations undertaken by a public body with a focus on children's well- being.	Oversight by a government agency focussing on education policy, with execution of the study contracted out to researchers.
Legislation	Use of existing legislations to facilitate data sharing for research purposes.	Enactment of dedicated legislation for mandating sharing of data to ContactPoint.	Use of existing legislations to operationalise a data ethics committee for facilitating data sharing for research purposes.	Adoption of legislation specifying the Commissioner's legal obligation to monitor well-being.	Children well-being legislation provided a facilitative environment for research.

Project	United Kingdom – Children's Well-being Measures	United Kingdom – ContactPoint	Australia – Children's Headline Indicators (CHI)	Western Australia – Well-being Monitoring Framework (WMF)	Alberta, Canada – Early Child Development Mapping Project (ECMap)
Type of data	Public access is available to 31 headline measures across seven domains, with measures compiled mainly from secondary data from existing studies and administrative sources.	Highly restricted use of basic identifying information with data contributed by national government, local authorities and practitioners.	Public access is available to 19 headline measures across three domains with data collection from studies/administrative sources from the public sector.	Online viewing of 123 headline measures across three domains with data collection from multiple sources is permissible.	Public access to aggregated research output with micro-level data gathered through primary data collection is made available.
Implementation consideration	Phased implementation with different modes of stakeholder engagement and use of guiding principles for development of measures.	Phased implementation with use of pilots, consultation with public and cost benefit assessment.	Phased implementation with consultation of state/territory government bodies, data-related committees and experts.	Phased implementation with collaboration with research institute for measures enhancement.	Execution of a statistical project with sustained momentum through proactive community engagement.

Project	United Kingdom – Children's Well-being Measures	United Kingdom – ContactPoint	Australia – Children's Headline Indicators (CHI)	Western Australia – Well-being Monitoring Framework (WMF)	Alberta, Canada – Early Child Development Mapping Project (ECMap)
Achievement or outcomes	Focus on raising awareness of specific children issues, contributing to advocacy work and substantiating an integrated narrative on children.	Closure due to concerns over privacy, inappropriate data access and cost overruns.	Use to inform priority areas of strategy and monitoring performance of children services.	Use to inform differential situation between general and vulnerable groups of children during the review of justice system for youths.	Collection of micro-level data enables improvement in service provision through identification of individual specific conditioning and risk factors.

Appendix C – List of Indicative Areas for Analysis

The table below sets out a list of potential areas for analysis in which stakeholders engaged have expressed interests. Stakeholders have also recognised the potential to conduct these analyses by developmental stage and of the two priority areas via the consolidation of existing data.

Potential areas for analysis	Description	Potential Indicator(s)
General children		
Number of diagnosed diseases of children in early childhood	Analyse aggregate health data regarding diagnosis of diseases, such as cancer, rare diseases and genetics disease, and new-born screening of children in early childhood to inform recommendations on health planning for early identification and treatment	• Prevalence of diseases among children in early childhood
Percentage of child obesity and associate health outcomes	Use existing aggregate data regarding children's health profiles to identify the obesity rate and associated health outcomes at different developmental stages	• Prevalence of child obesity and associated health outcomes among children of different age range
Correlation between children's health outcomes and living modes at each developmental stage	Analyse aggregate data regarding children's health profiles and their time spent on sleep, school, exercise and other activities (e.g. use of electronic devices) to identify any specific pattern of living mode and its impact on children's health across different development stages	• Symptoms/risk identification of potential adversarial living modes of children
Children with risk of abuse and neglect		
Statistically approximated unreported cases and number of undetected cases	Use existing aggregate data from SWD, HA and EDB and HKPF to identify unreported child abuse cases	• Prevalence of unreported child abuse cases

Potential areas for analysis	Description	Potential Indicator(s)
Percentage of reported case which required further investigation: • **Percentage of reported case investigated;** • **Percentage of reported case which required further investigation but not have been pursued with valid justifications; and** • **Percentage of reported case which required further investigation but have not been pursued due to other reasons**	Use existing aggregate data from SWD, HA, EDB and HKPF regarding the investigation status of reported child abuse cases	• Prevalence of child abuse cases which requires further investigation
Relationship of abuser(s) to the abused child(ren) and child(ren)'s immediate and indirect family members	Analyse existing aggregate data regarding the abusers and the abused children's immediate and indirect family profile to identify if there is any important red flags for observation	• Profiling of child abuser
General profile of abuser(s)	Analyse aggregate data from abusers' profiles (e.g. track record of substance abuse, employment status/nature, contextual factors, medical history, etc.) to identify commonalities among the abusers	• Profiling of child abuser
Physiological, emotional and socio-economic impacts of abuse or neglect on children	Analyse aggregate data from the abused/ neglected children's profiles and health outcomes to understand the impacts of abuse or neglect with a view to shed lights on identifying unreported children with risk of abuse or neglect or potential cases	• Symptom/ risk identification of potential children with risk of abuse or neglect

Potential areas for analysis	Description	Potential Indicator(s)
Children with SEN		
Number of diagnosed cases, statistically approximated number of undiagnosed cases and diagnosed cases at a delayed time	Collect and analyse existing aggregate data from HA and DH to identify the number of diagnosed SEN cases and compare it with the data from EDB to approximate the number of undiagnosed/ delayed diagnosed cases	• Prevalence of diagnosed and undiagnosed/ delayed SEN cases
Reasons for delayed diagnoses and undiagnosed cases	Collect and analyse existing aggregate data of the profiles of children with SEN diagnoses from HA/DH and compare them with the data from EDB to identify any discrepancies/inconsistencies in data/information collection to shed light on the potential reasons for delayed diagnoses and undiagnosed cases	• Profiling of delayed diagnoses and undiagnosed cases
Type of SEN diagnosed	Use existing aggregate data from relevant B/Ds (e.g. HA, EDB and DH) regarding the types of SEN diagnosed (and in comparison with overseas literature on SEN) to have a more comprehensive picture of SEN	• Profiling of SEN
Overall utilisation of support available (and estimated pent-up demand and waiting times)	Use existing aggregate data from HA, DH and EDB regarding support utilised by children with SEN (e.g. types of support, which type of support is in most demand, percentage of children with SEN receiving support, average waiting time for support, etc.)	• Prevalence/ frequency of utilisation of support available for SEN children

Potential areas for analysis	Description	Potential Indicator(s)
Outcomes of diagnosed cases upon different means of provision of support services (for example, private, public and collaboration between the two; on-time, and delayed support; sufficient and insufficient support)	Use existing aggregate data from HA, DH, and EDB to compare the profiles of different types of diagnosis cases (e.g. cases with appropriate support versus cases with insufficient support; cases with prompt support versus cases with delayed support; and case with support from private/ public/ both private and public organisations) to identify the impacts of each mean of support provision on children with SEN, such as understanding the trajectory of learning recovery of the different types of diagnosis cases.	• Effectiveness of different means of support provision

Appendix D – Examples of Data Raised by Stakeholders

Highlighting the importance of the domains of "health and safety" and "education and skills" in understanding children's development, stakeholders have raised the below examples of data pertaining to the two domains for inclusion into the scope of data.

Examples of data pertaining to the "health and safety" domain that could be considered include:

- Physical and mental health;
- Child mortality;
- Diseases and disorders;
- Child disabilities;
- Children with SEN (e.g. assessment and diagnosis);
- Child offenders;
- Abuse and family violence; and
- Social services regarding child protection

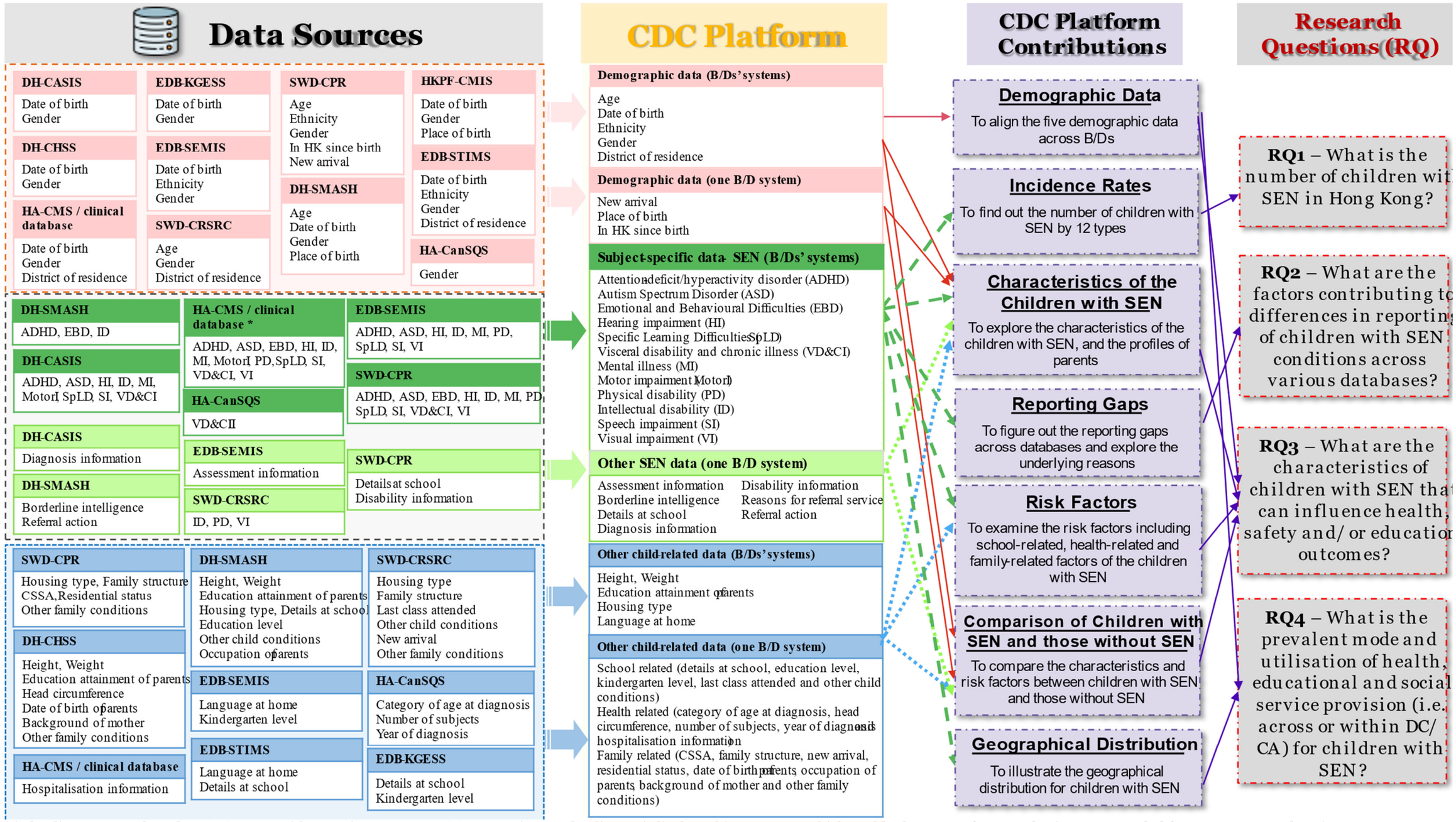
Examples of data pertaining to the "education and skills" domain that could be considered include:

- Academic data such as school grades and results;
- Dropout rate;
- Language spoken at home;
- Language proficiency;
- Children encountering language barriers; and
- Early childhood development, such as educational performance of children with SEN, especially kindergarten students.

Appendix E - Anticipated Relations and Contributions of CDC Platform Upon Data Alignment

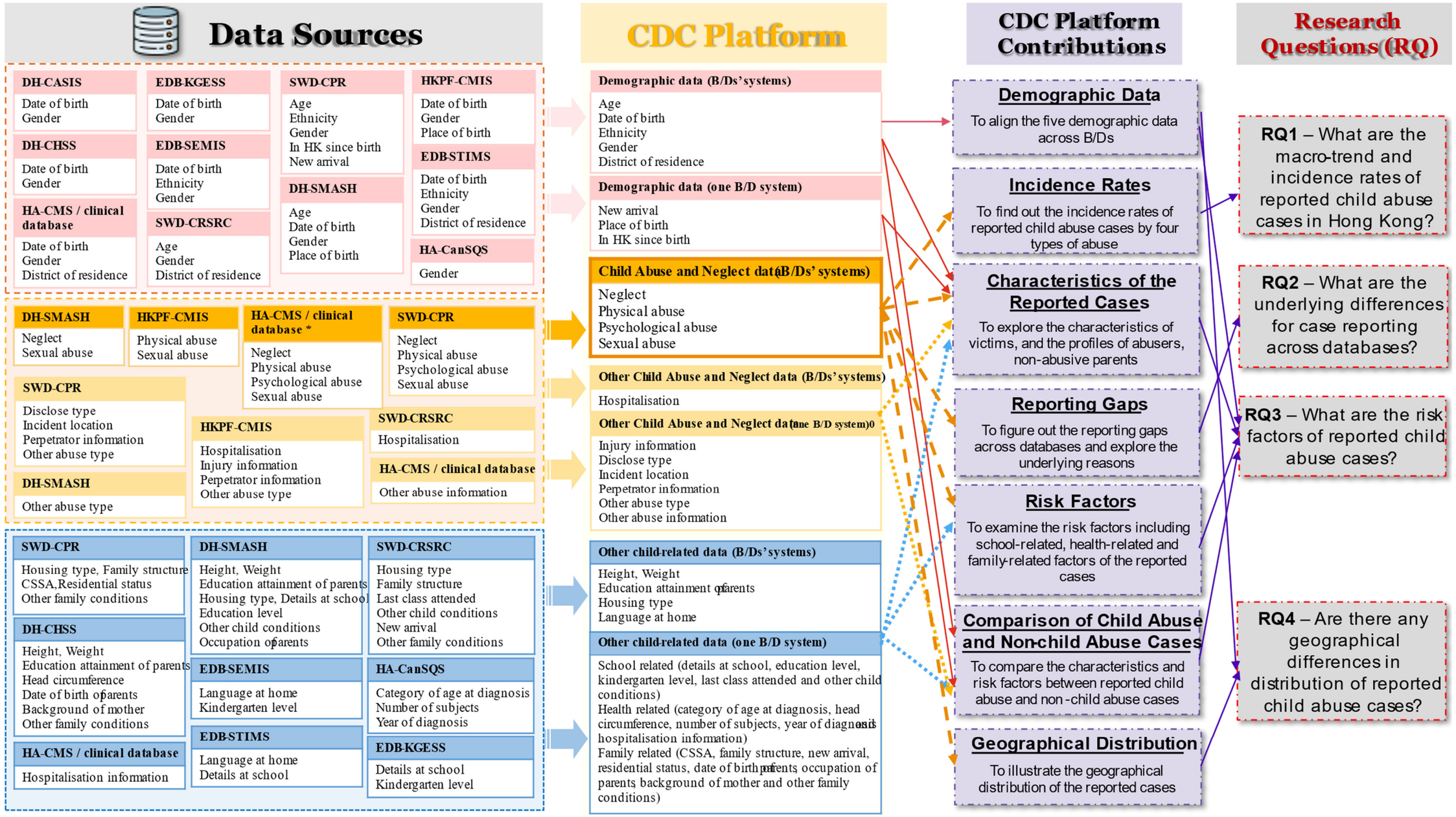
The figure on the following pages illustrates (i) the anticipated relations between data sources and the CDC platform; (ii) the corresponding contributions of the CDC platform on relevant demographic data, subject specific data and other child-related data; and (iii) how these contributions could inform some of the illustrative research questions set out in **Section 2.1.1.1.3** , based on the data alignment activities set out in **Section 2.1.2.2**.

Anticipated Relations and Contributions of CDC Platform Upon Data Alignment



(*) the diagnoses are based on ICD-9-CM,with mapping to 2010 v ICD-10. Review on the diagnosis list by subject experts, which could take up to a few months, is recommended for a more comprehensive diagnosis list for child abuse & neglect, and SEN. The number of variables in the data category are therefore to be determined after the expert review.

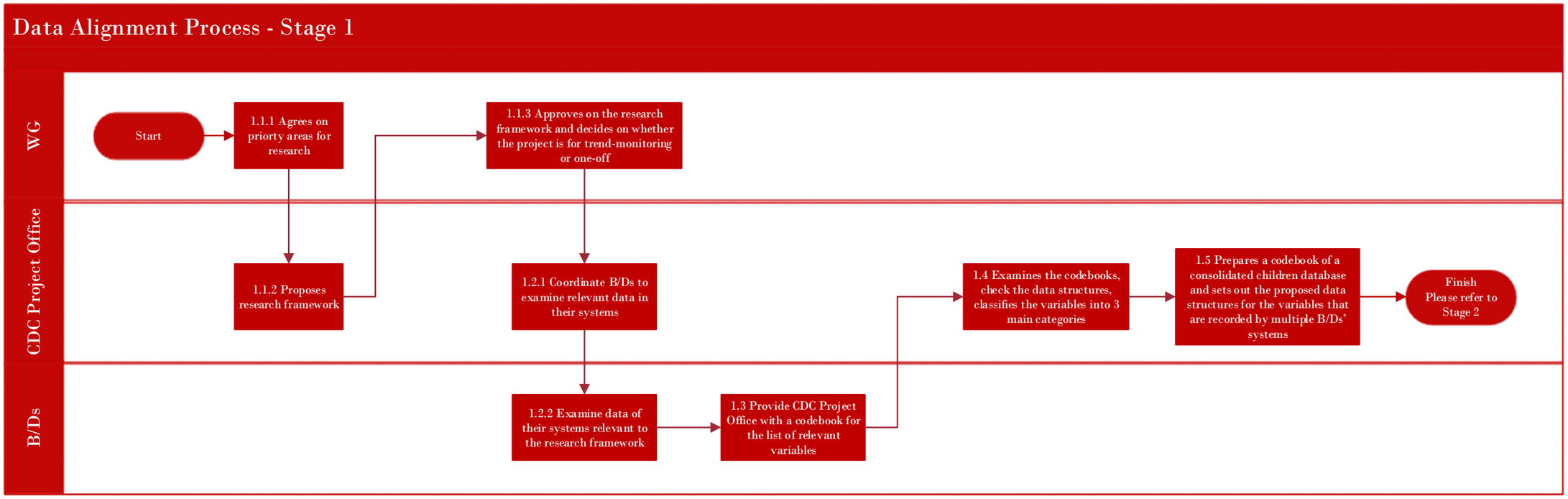
(Cont'd) Anticipated Relations and Contributions of CDC Platform Upon Data Alignment

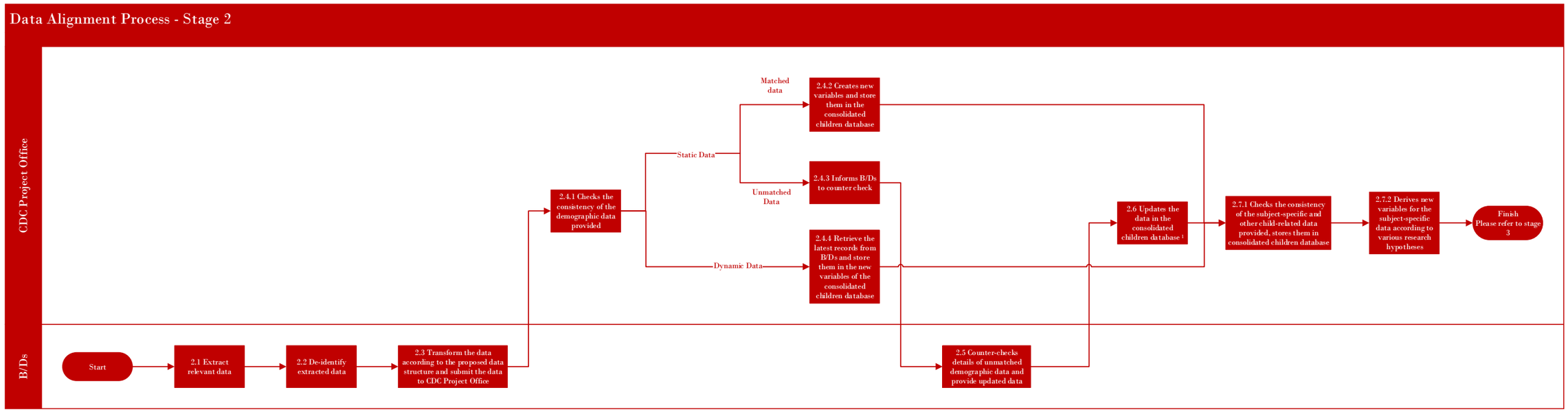


(*) the diagnoses are based on ICD-9-CM,with mapping to 2010 v ICD-10. Review on the diagnosis list by subject experts, which could take up to a few months, is recommended for a more comprehensive diagnosis list for child abuse & neglect, and SEN. The number of variables in the data category are therefore to be determined after the expert review.

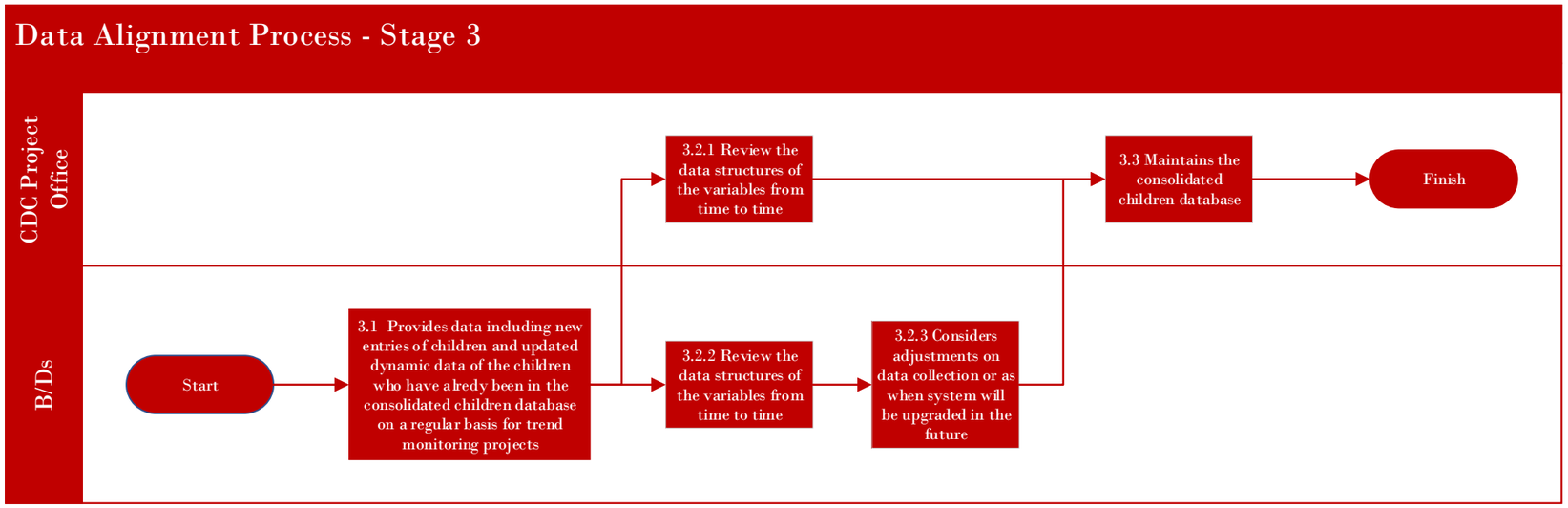
Appendix F - Process Diagram for Three Stages of Data Alignment

The figures below illustrates the process involved in the Three Stages of Data Alignment proposed in **Section 2.1.2.3.2.**





Note: Step 2.4 to Step 2.6 should be repeated until the demographic static data are matched or in some circumstances such as unavailable information and incomplete checking, consensuses have been made across B/Ds on the inconsistent demographic static data to be recorded in the consolidated children database.


Data Alignment Process - Stage 3
CDC Project Office
B/Ds
3.2.1 Review the data structures of the variables from time to time
3.3 Maintains the consolidated children database
Finish
Start
3.1 Provides data including new entries of children and updated dynamic data of the children who have alredy been in the consolidated children database on a regular basis for trend monitoring projects
3.2.2 Review the data structures of the variables from time to time
3.2.3 Considers adjustments on data collection or as when system will be upgraded in the future

Appendix G - List of Proposed Software for Data Analysis

Software for Data Analysis
SAS Decision Manager (including SAS Data Integration Server)
SAS Viya • SAS Visual Analytics • SAS Visual Statistics • SAS Visual Data Mining and Machine Learning
Apache Hadoop
MySQL or MariaDB
MongoDB

Software for Data Dissemination
SAS Enterprise Guide (Optional)
Programming Tools: • R • Python • Open-source Machine Learning Library

Appendix H – Summary of Current State of Relevant Databases of External Parties

Database Name		CDI	KeySteps@ JC	YCDI	1997 Birth Cohort Study
Start Date		2006	2018	2014	1997
End Date/Anticipated End Date		N/a	2022 [25]	N/a	N/a
Development Objectives: Current focus areas and benefits					
Focus Area(s)	Monitoring macrotrends	✓	✓	✓	✓
	Monitoring performance related to programmes/ services provided to children		✓	✓	
	Understanding outcome of interventions on general children/ specific segment of children		✓	✓	
	For compliance of international reporting protocols or research purposes		✓	✓	
	Others				✓ [26]
Benefit(s)	Ease of monitoring the impact/outcome of programme/services provided		✓	✓	
	Ease of monitoring macro-trends	✓	✓	✓	
	More informed decision-making by senior management/policy makers	✓	✓		
	Better programme/scheme/service design	✓	✓	✓	
	Support social or health research	✓	✓	✓	✓

[25] It is a five-year project anticipated to end in 2022. However, HKJC plans to continue the project and is currently reviewing its data strategy. The KeySteps@JC project may potentially be renewed with some modifications.

[26] It refers to exploring determinants of health (e.g genetics, behavioural, anthropometrics, and biomarkers) in a non-Western setting, which may shed light in the understanding of drivers of diseases.

Database Name		CDI	KeySteps@ JC	YCDI	1997 Birth Cohort Study
	Better collaboration across sectors, e.g. public, private, NGO involvement		✓		
Scope of Data: Type and Domain of Data					
Domain(s)	Health and Safety	✓	✓	✓	✓
	Education and Skills	✓	✓	✓	
	Material well-being	✓	✓	✓	✓
	Family and Peer relationship		✓	✓ (Family)	
	Behaviours and Risks	✓	✓	✓	
	Others	✓ [27]	✓	✓ (Population)	✓ [28]
General or Targeted Segment(s) [29]	General (Non-Targeted) [30]	✓	✓	✓	✓
	Ethnic Minorities			✓	
	Special Education Needs			✓	
	Chronic Health Condition [31]			✓ [32]	
	Risk of Neglect / Abuse			✓	
	Poverty		✓	✓	

[27] This refers to children demographic, family background, and environmental factors.
[28] This is a cohort study exploring determinants of health, not only restricted to children but also extends to adulthood upon continual follow up of the participants from birth in 1997.
[29] These segments were previously indicated in the preliminary proposal of the Initial Work Plan of the Commission on Children (CoC) as proposed by the Preparatory Committee chaired by the Chief Executive in a discussion paper for establishing the CoC on 27 February 2018.
[30] General (Non-targeted) refers to a case whereby the database includes data of children generally without specifically segmenting data further by any of the target segments indicated in the table.
[31] The term "Acute Health Conditions" used in the questionnaire distributed for local database stocktaking exercise was adjusted to "Chronic Health Conditions" based on our engagement with stakeholders and suggestions from representative from the healthcare sector. With reference to a Child Health Survey published by CHP, Chronic Health Conditions include illnesses such as visual problems, allergic rhinitis, eczema, food allergy and asthma etc.
[32] To be verified

Database Name		CDI	KeySteps@JC	YCDI	1997 Birth Cohort Study
	Special Family Background			✓	
Age Segment(s)	General Children	Age 0-18 years		Age below 12	
	Specific Segments of Children		Age 3-6 years (Children studying in kindergarten in Sham Shui Po and Tin Shui Wai that were selected as pilot schools[33])	Age below 12	
	Adults		Age >18 years		Born in April and May, 1997
Level of Details Collected	Individual		✓		✓
	Household		✓		
	District/Regional	✓	✓	✓	
Identifier			✓		✓
Data Architecture					
Database Management System (DBMS)	Single Tier[34]			✓	
	Two Tiers[35]				✓
	Three Tiers[36]		✓		
	Not Applicable	✓			

[33] Sham Shui Po and Tin Shui Wai were selected as they are districts with a higher poverty rate.

[34] "Single tier" is defined as a DBMS whereby the presentation, application, and database are all in same layer, such as standalone desktop application running on dedicated computer workstations, e.g. Excel or Access Database

[35] "Two tiers" is defined as a DBMS whereby there are separate presentation and database layers, such as multi-user computer application running on individual workstations with corresponding data files residing on shared network drive, e.g. End-user Computing Systems

[36] "Three tiers" is defined as a DMBS with separate presentation, application, and database Layers, such as web-based application with business logic and database resided on separate computer servers.

Database Name		CDI	KeySteps@ JC	YCDI	1997 Birth Cohort Study
Frequency of update	Regular		✓	✓	
	Ad-hoc[37]	✓	✓		✓
	Non-recurrent				
Availability of Data Dictionary[38]			✓		✓
User Group(s)	General Public	✓		✓	
	Other specific organisations[39]		✓ (HKU)		
	Other Government bodies				
	Business Users of the database within your organisation		✓		
	IT/System Team in your organisation				
	Other Authorised Persons in your organisation[40]		✓		✓
Data Security and Classification	Unclassified	✓		✓	
	Restricted		✓		✓
	Confidential		✓		✓
	Secret				
	Top Secret				
Current Data Sharing Arrangement(s)	Individual		✓		✓
	Aggregated		✓	✓	
Potential for Data Sharing in the Future[41]	With Other Government B/Ds	NO	AGGR	AGGR	AGGR
	With NGOs	NO	AGGR	AGGR	AGGR
	With Academia	NO	BOTH	AGGR	BOTH

[37] This refers to situations where data may be updated more than once but not on a regular basis.
[38] This refers to the availability of System Design Document documenting data fields stored in the database (e.g. Data Manual, Entity Relationship Diagram, Data Flow Diagram, Data Matrix, Data Mapping Specification, Data Dictionary, etc.)
[39] This refers to organisations other than Government entities. Examples include academic institutions or NGOs.
[40] This refers to other authorised personnel (other than business users or IT support staff) within the organisation who are granted access to the stated database. Examples include administrative users and clerical officers.
[41] There are four options: (i) AGGR: Aggregated Level only; (ii) IND: Individual Level only; (iii) BOTH: Aggregated and Individual level; (iv) NO: not open for future data sharing opportunities.

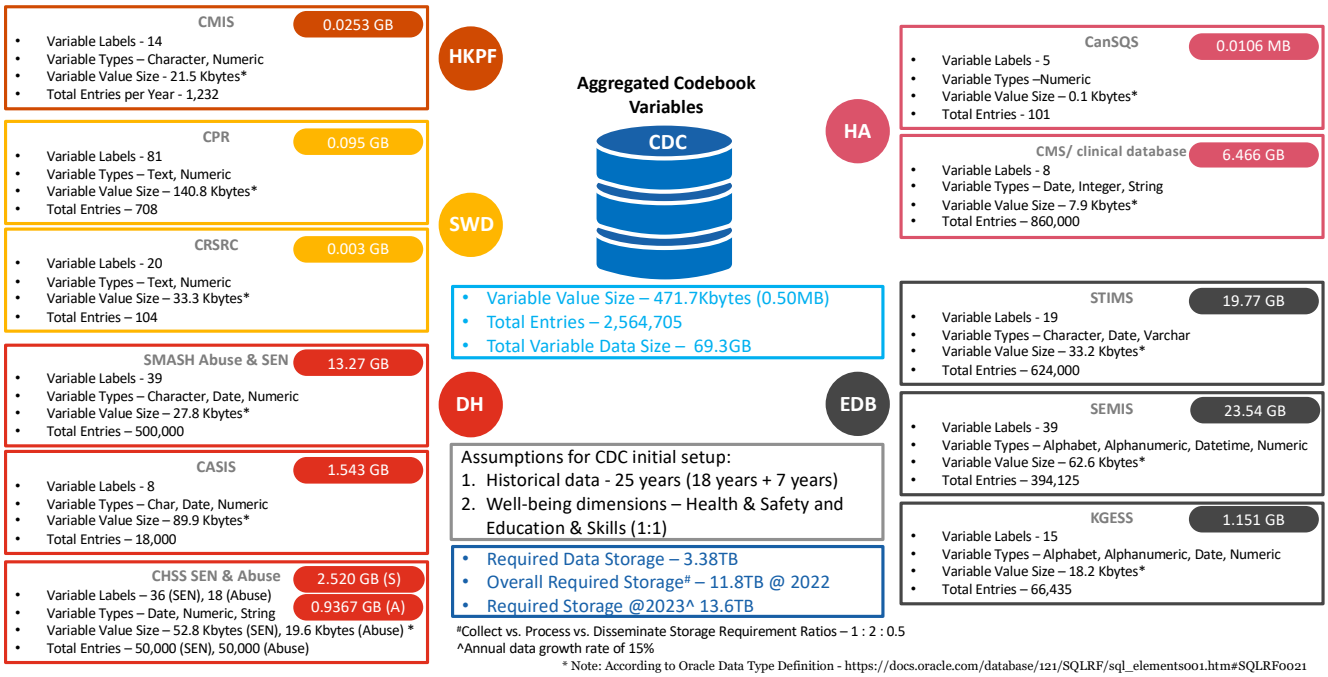
Database Name		CDI	KeySteps@ JC	YCDI	1997 Birth Cohort Study
	With Private Sector	NO	NO	AGGR	NO
	With General Public	NO	NO	AGGR	AGGR

Appendix I – Approach and Assumptions for Sizing Estimation

- The 0.50MB aggregated data size is based on the consolidated data fields from the initial codebooks provided by the five B/Ds, and the associated data types for field size estimation with reference to typical RDBMS, in this case Oracle. This represents a single aggregated entry (row) of data in CDC to facilitate the data analysis covering "Children with Risk of Abuse and Neglect" and "Children with SEN".
- There are a total number of 2,564,705 entries of data in total per year from the eleven systems examined. Thus, the aggregated data field sizes from all the provided codebooks totalled to 69.3GB.
- Since the aggregated data only accounted for one year, and the fact that CDC needs to maintain at least 18 years of data covering individuals from infancy to adulthood, and subsequently another 7 years according to the recordkeeping practices of the Government Record Service, i.e. 18+7, a total storage of 25 years of data should be allocated during the initial setup of the CDC, resulting in 1.69TB of storage requirement.
- For Foundational Mode, the data is expected to cover both "Health & Safety" and "Education & Skills" children well-being dimensions. Since the provided codebooks only cover "Health & Safety", with an assumed ratio of 1:1, another 1.69TB is allocated for "Education & Skills". Thus a total of 3.38TB for collected data as required in 2022.
- For Enhancement Mode, the data is expected to cover all five children well-being dimensions, namely, "Health & Safety", "Education & Skills", "Material Well-being", "Behaviours & Risks" and "Family & Peer Relationships". Since the provided codebooks only cover "Health & Safety", with an assumed ratio of 1:1 for "Education & Skills", another 1.69 TB is allocated. For the remaining three dimensions, it is assumed a ration distribution of 1/3 : 1/3 : 1/3 ratio with respect to "Health & Safety". Therefore, another 0.56 TB is allocated to either of the remaining three dimensions respectively, resulting a total of 5.06TB for collected data as required in 2022.
- Figures below illustrate the number of entries per year of individual systems and how the data sizing requirement is derived from the aggregated data size of 0.50MB.

Deriving the sizing requirement for Foundational Mode

CDC Data Sizing Estimation – Foundational Mode



#Collect vs. Process vs. Disseminate Storage Requirement Ratios – 1 : 2 : 0.5
^Annual data growth rate of 15%

* Note: According to Oracle Data Type Definition - https://docs.oracle.com/database/121/SQLRF/sql_elements001.htm#SQLRF0021

Aggregated Cookbook Variables	2022 (TB)	Priority Areas	Child Well-being Dimensions
From HKPF, SWD, DH, EDB & HA	1.69 (69.32GB x 25 Years)	SEN and Child Abuse	Health & Safety

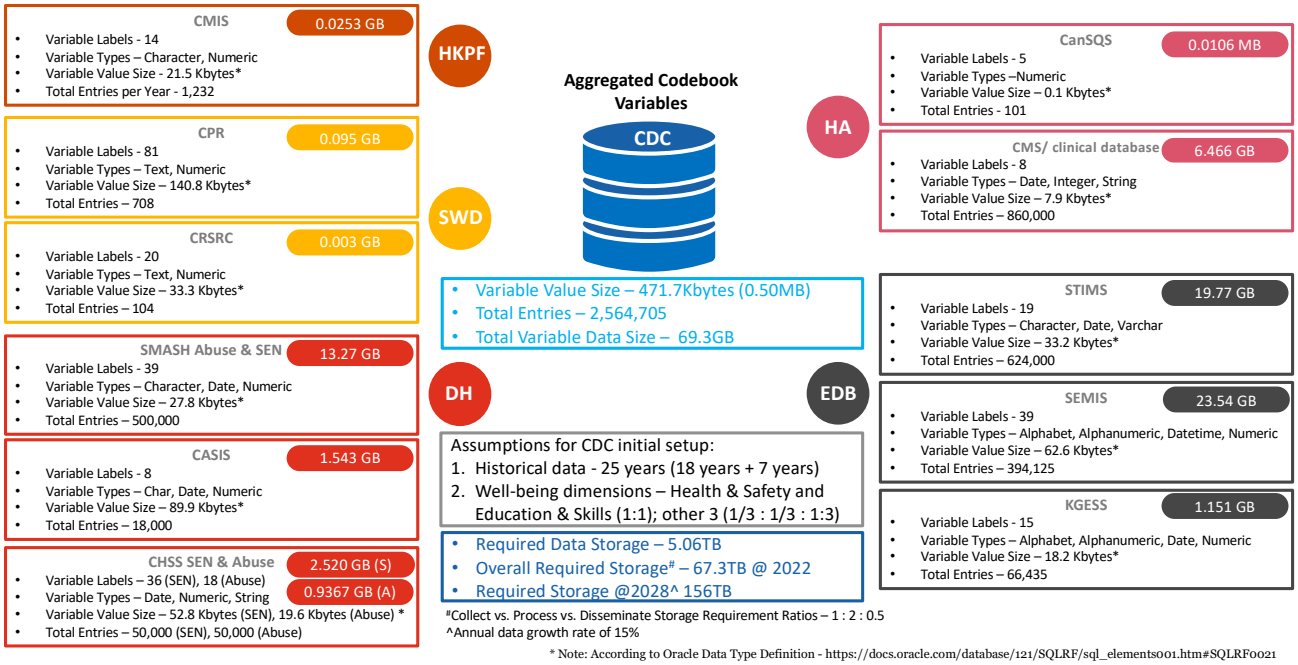
CDC Scope of Data for Foundational Mode	2022 (TB)	Ratio	Storage (TB)
1. Health & Safety	1.69	1	1.69
2. Education & Skills	(assume same as #1)	1	1.69
3. Material Well-being	N/A	0	0
4. Behaviours & Risks	N/A	0	0
5. Family & Peer Relationships	N/A	0	0
		Total:	**3.38**

Projected Storage Requirements	2022 (TB)	Ratio	Storage (TB)
Data Collection	3.38	1	3.38
Data Processing and Analysis	-	2	6.77
Data Dissemination	-	0.5	1.69
		Total:	11.85

Annual Growth Rate	Year 1 - 2023 (TB)	Year 2 – 2024 (TB)	Year 3 - 2025 (TB)	Year 4 - 2026 (TB)	Year 5 - 2027 (TB)
15%	13.6	15.7	18.0	20.7	23.8

Deriving the sizing requirement for Enhancement Mode

CDC Data Sizing Estimation – Enhancement Mode



#Collect vs. Process vs. Disseminate Storage Requirement Ratios – 1 : 2 : 0.5
^Annual data growth rate of 15%

* Note: According to Oracle Data Type Definition - https://docs.oracle.com/database/121/SQLRF/sql_elements001.htm#SQLRF0021

Aggregated Cookbook Variables	2022 (TB)	Priority Areas	Child Well-being Dimensions
From HKPF, SWD, DH, EDB & HA	1.69 (69.32GB x 25 Years)	SEN and Child Abuse	Health & Safety

CDC Scope of Data for Enhancement Mode	2022 (TB)	Ratio	Storage (TB)
1 . Health & Safety	1.69	1	1.69
2. Education & Skills	(assume same as #1)	1	1.69
3. Material Well-being	(assume 1/3 as #1)	1/3	0.56
4. Behaviours & Risks	(assume 1/3as #1)	1/3	0.56
5. Family & Peer Relationships	(assume 1/3 as #1)	1/3	0.56
		Total (for 4 B/Ds and one statutory body) :	**5.06**
Extrapolated to cover involving B/D, statutory bodies, NGO, Academia interviewed (Total 19 bodies)*:			**19.23**

Projected Storage Requirements	2022 (TB)	Ratio	Storage (TB)
Data Collection	19.23	1	19.23
Data Processing and Analysis	-	2	38.46
Data Dissemination	-	0.5	9.62
		Total:	67.31

* 10 B/Ds and statutory bodies including: EDB, DH, HA, SWD, HKPF, HAD, C&SD, DOJ, Housing Authority, Judiciary – Juvenile Court; 4 NGOs including: Boy's & Girl's Clubs Association, Hong Kong Society for the Protection of Children, Hong Kong Council of Social Services, Hong Kong Jockey Club; 5 Academia including School of Public Health of HKU, Department of Paediatrics of CU, Faculty of Law of HKU, Faculty of Law of CU and Advisory Committee on Mental Health

4

Storage of relevant children data from infancy to adulthood (i.e., 18 years plus 7 years for data retention)

Annual Growth Rate	2023 (TB)	2024 (TB)	2025 (TB)	2026 (TB)	2027 (TB)
15%	77	89	102	118	135
Annual Growth Rate	**Year 1 - 2028 (TB)**	**Year 2 – 2029 (TB)**	**Year 3 - 2030 (TB)**	**Year 4 – 2031 (TB)**	**Year 5 - 2032 (TB)**
15%	156	179	206	237	272
Annual Growth Rate	**Year 6 - 2033 (TB)**	**Year 7 – 2034 (TB)**	**Year 8 - 2035 (TB)**	**Year 9 - 2036 (TB)**	**Year 10 - 2037 (TB)**
15%	313	360	414	476	548
Annual Growth Rate	**Year 11 - 2038 (TB)**	**Year 12 – 2039 (TB)**	**Year 13 - 2040 (TB)**	**Year 14 - 2041 (TB)**	**Year 15 - 2042 (TB)**
15%	630	724	833	958	1,101
Annual Growth Rate	**Year 16 - 2043 (TB)**	**Year 17 – 2044 (TB)**	**Year 18 - 2045 (TB)**	**Year 19 - 2046 (TB)**	**Year 20 - 2047 (TB)**
15%	1,267	1,457	1,675	1,926	2,215

Appendix J – Case Illustration – Data Linkage Study on Child Abuse

Objectives

A study[42] was conducted to examine the association between health problems and child abuse by linking SWD and HA datasets. In particular, the study leveraged two major Government databases on child abuse in Hong Kong:

- **SWD:** Child Protection Registry (CPR); and
- **HA:** Clinical Data Analysis and Reporting System (CDARS) and Accident and Emergency Information System (AEIS).

Outcomes

Table 6. Association between child abuse and health problems.

Any diagnosis (ICD-10)		Prevalence among child abuse victims	Prevalence in H.K. population[1]	Odds ratio (95% CI)	p-value
X60 to X84	History of suicidal attempt	1.92%	0.02%	**96.60 (80.04, 116.00)**	<0.0001
X60 to X84/ S00 to T98	History of Injury	23.9%[2]	3.2%	**9.48 (8.13, 11.03)**	<0.0001
F00 to F99	Mental health problems	10.8%	1.2%	**9.97 (9.28, 10.71)**	<0.0001
Q00 to Q99	Congenital Malformations/ Chromosomal Abnormalities	4.0%	1.3%	**3.17 (2.82, 3.54)**	<0.0001

[1] Prevalence among children in H.K. was estimated using HA CDARS
[2] Injury prevalence of child abuse victims was represented by SWD only group

[42] https://www.pico.gov.hk/doc/en/research_reports/CPU_research_report-epidemiology_of_child_abuse_and_its_geographic_distribution_in_hong_kong.pdf

Findings suggest a higher risk of key health problems (i.e. suicidal attempt, injury, mental health problems, and congenital malformations, chromosomal abnormalities) experienced by victims of child abuse. Specifically:

- Risk of suicide attempt was higher among children suffering from sexual abuse, psychological abuse and children experiencing multiple abuses;
- Mental health diagnoses were more common in victims of psychological and multiple abuses; and
- Congenital malformations and chromosomal abnormalities were more commonly found amongst neglected children.

Implication for policymaking and service planning

Such finding could serve as evidence for the need of improvement on the case management system by involving both social services and medical professionals to provide early intervention and to address the needs of the abused children suffering from mental health problems and to reduce attempted suicide rate amongst victims of child abuse.

Commercial-in-Confidence

This report has been prepared for, and only for, the Labour and Welfare Bureau ("LWB") of the Government of Hong Kong Special Administrative Region in accordance with the terms of the Letter of Acceptance (issued by LWB on 29 November 2019) and of our proposal, and for no other purpose. We do not accept or assume any liability or duty of care for any other purpose or to any other person to whom this report is shown or into whose hands it may come save where expressly agreed by our prior consent in writing.